2023 • 27 • 7

Print ISSN 2500-0462
Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.
Периодичность 8 выпусков в год
DOI 10.18699/VJGB-23-83

Учредители

Сибирское отделение Российской академии наук

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Межрегиональная общественная организация Вавиловское общество генетиков и селекционеров

2023 • 27 • 7

Print ISSN 2500-0462
Online ISSN 2500-3259

Online edition

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

Founded in 1997
Published 8 times annually
DOI 10.18699/VJGB-23-83

Founders

Siberian Branch of the Russian Academy of Sciences

Federal Research Center Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

СОДЕРЖАНИЕ • 2023 • 27 • 7

725 ОТ РЕДАКТОРА
Н.А. Колчанов, Ю.Г. Матушкин

Компьютерная геномика

728 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования. *С.В. Филонов, Н.Л. Подколодный, О.А. Подколодная, Н.Н. Твердохлеб, П.М. Пономаренко, Д.А. Рассказов, А.Г. Богомолов, М.П. Пономаренко*

737 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования. *А.Ю. Пронозин, Е.А. Салина, Д.А. Афонников*

Системная компьютерная биология

746 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Центральный регуляторный контур генной сети морфогенеза механорецепторов дрозофилы: анализ *in silico*. *Т.А. Бухарина, В.П. Голубятников, Д.П. Фурман*

755 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Бифуркационный анализ мультистабильности и гистерезиса в модели ВИЧ-инфекции. *И.В. Миронов, М.Ю. Христиченко, Ю.М. Нечепуренко, Д.С. Гребенников, Г.А. Бочаров*

768 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Применение генных сетей к анализу результатов метаболомного скрининга плазмы крови пациентов с послеоперационным делирием. *В.А. Иванисенко, Н.В. Басов, А.А. Макарова, А.С. Вензель, А.Д. Рогачев, П.С. Деменков, Т.В. Иванисенко, М.А. Клещев, Е.В. Гайслер, Г.Б. Мороз, В.В. Плеско, Ю.С. Сотникова, Ю.В. Патрушев, В.В. Ломиворотов, Н.А. Колчанов, А.Г. Покровский*

776 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Молекулярно-генетические пути регуляции вирусом гепатита С экспрессии клеточных факторов PREB и PLA2G4C, играющих важную роль для репликации вируса. *Е.Л. Мищенко, А.А. Макарова, Е.А. Антропова, А.С. Вензель, Т.В. Иванисенко, П.С. Деменков, В.А. Иванисенко*

784 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Приоритизация потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза, на основе реконструкции и анализа ассоциативных генных сетей. *П.С. Деменков, Е.А. Антропова, А.В. Адамовская, Е.Л. Мищенко, Т.М. Хлебодарова, Т.В. Иванисенко, Н.В. Иванисенко, А.С. Вензель, И.Н. Лаврик, В.А. Иванисенко*

794 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
База знаний RatDEGdb по дифференциально экспрессирующимся генам крысы как модельного объекта биомедицинских исследований. *И.В. Чадаева, С.В. Филонов, К.А. Золотарева, Б.М. Хандаев, Н.И. Ершов, Н.Л. Подколодный, Р.В. Кожемякина, Д.А. Рассказов, А.Г. Богомолов, Е.Ю. Кондратюк, Н.В. Климова, С.Г. Шихевич, М.А. Рязанова, Л.А. Федосеева, О.Е. Редина, О.С. Кожевникова, Н.А. Стефанова, Н.Г. Колосова, А.Л. Маркель, М.П. Пономаренко, Д.Ю. Ощепков*

Структурная компьютерная биология и фармакология

807 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Применение метода взвешенных гистограмм для расчета термодинамических параметров формирования комплексов олигодезоксирибонуклеотидов. *И.И. Юшин, В.М. Голышев, Д.В. Пышный, А.А. Ломзов*

Эволюционная компьютерная биология

815 ОБЗОР
Внутриопухолевая гетерогенность: модели возникновения и эволюции злокачественных опухолей. *Р.А. Иванов, С.А. Лашин*

820 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Поиск дифференциально метилированных регионов в геномах древних и современных людей. *Д.Д. Бородко, С.В. Женило, Ф.С. Шарко*

829 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Анализ особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, на основе индексов филостратиграфического возраста и микроэволюционной изменчивости. *Е.В. Игнатьева, С.А. Лашин, З.С. Мустафин, Н.А. Колчанов*

839 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
О пространстве вариантов генетических последовательностей SARS-CoV-2. *А.Ю. Пальянов, Н.В. Пальянова*

Методы глубокого машинного обучения для биоинформатики и системной биологии

851 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Сверточные нейронные сети для классификации по данным ЭЭГ здоровых людей, практикующих или не практикующих медитацию.
С. Фу, С.С. Таможников, А.Е. Сапрыгин, Н.А. Истомина, Д.И. Клемешова, А.Н. Савостьянов

859 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Определение содержания меланина и антоцианов в зернах ячменя на основе анализа цифровых изображений методами машинного обучения.
Е.Г. Комышев, М.А. Генаев, И.Д. Бусов, М.В. Кожекин, Н.В. Артеменко, А.Ю. Глаголева, В.С. Коваль, Д.А. Афонников

Экологическая компьютерная биология

869 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Математическое моделирование динамики кворум-эффекта в накопительной культуре люминесцентных бактерий *Photobacterium phosphoreum* 1889.
С.И. Барцев, А.Б. Сарангова

878 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Математическая модель системы жизнеобеспечения на основе водорослей, замкнутая по кислороду и углекислому газу. *Д.А. Семёнов, А.Г. Дегерменджи*

884 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Феноменологическая модель негеномной изменчивости люминесцентных бактериальных клеток. *С.И. Барцев (на англ. языке)*

Компьютерная биология растений

890 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
DyCeModel: программное средство для одномерного моделирования распределения гормонов растений, контролирующих образование структуры ткани. *Д.С. Азарова, Н.А. Омельянчук, В.В. Миронова, Е.В. Землянская, В.В. Лавреха (на англ. языке)*

Техническая биоинформатика

898 ОБЗОР
Лабораторные информационные системы для управления исследовательскими работами в биологии. *А.М. Мухин, Ф.В. Казанцев, С.А. Лашин*

Молекулярная и клеточная биология

906 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Анализ транскрипционной активности модельных piggyBac-трансгенов, стабильно интегрированных в разные локусы генома культивируемых клеток CHO при отсутствии селекционного давления. *Л.А. Яринич, А.А. Огиенко, А.В. Пиндюрин, Е.С. Омелина*

VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2023 • 27 • 7

725 FROM THE EDITOR
N.A. Kolchanov, Yu.G. Matushkin

Computational genetics

728 ORIGINAL ARTICLE
Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *S.V. Filonov, N.L. Podkolodnyy, O.A. Podkolodnaya, N.N. Tverdokhleb, P.M. Ponomarenko, D.A. Rasskazov, A.G. Bogomolov, M.P. Ponomarenko*

737 ORIGINAL ARTICLE
GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing. *A.Y. Pronozin, E.A. Salina, D.A. Afonnikov*

Systems computational biology

746 ORIGINAL ARTICLE
The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis. *T.A. Bukharina, V.P. Golubyatnikov, D.P. Furman*

755 ORIGINAL ARTICLE
Bifurcation analysis of multistability and hysteresis in a model of HIV infection. *I.V. Mironov, M.Yu. Khristichenko, Yu.M. Nechepurenko, D.S. Grebennikov, G.A. Bocharov*

768 ORIGINAL ARTICLE
Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium. *V.A. Ivanisenko, N.V. Basov, A.A. Makarova, A.S. Venzel, A.D. Rogachev, P.S. Demenkov, T.V. Ivanisenko, M.A. Kleshchev, E.V. Gaisler, G.B. Moroz, V.V. Plesko, Y.S. Sotnikova, Y.V. Patrushev, V.V. Lomivorotov, N.A. Kolchanov, A.G. Pokrovsky*

776 ORIGINAL ARTICLE
Molecular-genetic pathways of hepatitis C virus regulation of the expression of cellular factors PREB and PLA2G4C, which play an important role in virus replication. *E.L. Mishchenko, A.A. Makarova, E.A. Antropova, A.S. Venzel, T.V. Ivanisenko, P.S. Demenkov, V.A. Ivanisenko*

784 ORIGINAL ARTICLE
Prioritization of potential pharmacological targets for the development of anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway: the reconstruction and analysis of associative gene networks help. *P.S. Demenkov, E.A. Antropova, A.V. Adamovskaya, E.L. Mishchenko, T.M. Khlebodarova, T.V. Ivanisenko, N.V. Ivanisenko, A.S. Venzel, I.N. Lavrik, V.A. Ivanisenko*

794 ORIGINAL ARTICLE
RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research. *I.V. Chadaeva, S.V. Filonov, K.A. Zolotareva, B.M. Khandaev, N.I. Ershov, N.L. Podkolodnyy, R.V. Kozhemyakina, D.A. Rasskazov, A.G. Bogomolov, E.Yu. Kondratyuk, N.V. Klimova, S.G. Shikhevich, M.A. Ryazanova, L.A. Fedoseeva, O.E. Redina, O.S. Kozhevnikova, N.A. Stefanova, N.G. Kolosova, A.L. Markel, M.P. Ponomarenko, D.Yu. Oshchepkov*

Structural computational biology and pharmacology

807 ORIGINAL ARTICLE
Application of the weighted histogram method for calculating the thermodynamic parameters of the formation of oligodeoxyribonucleotide duplexes. *I.I. Yushin, V.M. Golyshev, D.V. Pyshnyi, A.A. Lomzov*

Evolutionary computational biology

815 REVIEW
Intratumor heterogeneity: models of malignancy emergence and evolution. *R.A. Ivanov, S.A. Lashin*

820 ORIGINAL ARTICLE
Search for differentially methylated regions in ancient and modern genomes. *D.D. Borodko, S.V. Zhenilo, F.S. Sharko*

829 ORIGINAL ARTICLE
Evolution of human genes encoding cell surface receptors involved in the regulation of appetite: an analysis based on the phylostratigraphic age and divergence indexes. *E.V. Ignatieva, S.A. Lashin, Z.S. Mustafin, N.A. Kolchanov*

839 ORIGINAL ARTICLE
On the space of SARS-CoV-2 genetic sequence variants. *A.Yu. Palyanov, N.V. Palyanova*

Deep learning methods in bioinformatics and systems biology

851 ORIGINAL ARTICLE
Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data. *X. Fu, S.S. Tamozhnikov, A.E. Saprygin, N.A. Istomina, D.I. Klemeshova, A.N. Savostyanov*

859 ORIGINAL ARTICLE
Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods. *E.G. Komyshev, M.A. Genaev, I.D. Busov, M.V. Kozhekin, N.V. Artemenko, A.Y. Glagoleva, V.S. Koval, D.A. Afonnikov*

Ecological computational biology

869 ORIGINAL ARTICLE
Mathematical modeling of quorum sensing dynamics in batch culture of luminescent bacterium *Photobacterium phosphoreum* 1889. *S.I. Bartsev, A.B. Sarangova*

878 ORIGINAL ARTICLE
Alga-based mathematical model of a life support system closed in oxygen and carbon dioxide. *D.A. Semyonov, A.G. Degermendzhi*

884 ORIGINAL ARTICLE
A phenomenological model of non-genomic variability of luminescent bacterial cells. *S.I. Bartsev*

Computational plant biology

890 ORIGINAL ARTICLE
DyCeModel: a tool for 1D simulation for distribution of plant hormones controlling tissue patterning. *D.S. Azarova, N.A. Omelyanchuk, V.V. Mironova, E.V. Zemlyanskaya, V.V. Lavrekha*

Industrial bioinformatics

898 REVIEW
Laboratory information systems for research management in biology. *A.M. Mukhin, F.V. Kazantsev, S.A. Lashin*

Molecular and cell biology

906 ORIGINAL ARTICLE
Analysis of the transcriptional activity of model piggyBac transgenes stably integrated into different loci of the genome of CHO cells in the absence of selection pressure. *L.A. Yarinich, A.A. Ogienko, A.V. Pindyurin, E.S. Omelina*

Вавиловский журнал генетики и селекции. 2023;27(7):725-727
DOI 10.18699/VJGB-23-84

N.A. Kolchanov

Yu.G. Matushkin

Dear colleagues, dear readers!

We present to your attention a new issue of the Vavilov Journal of Genetics and Breeding dedicated to bioinformatics and systems computational biology. These areas of scientific research are now rapidly transforming as natural sciences enter the era of big data. Intense development of the omics technologies (genomics, transcriptomics, proteomics, metabolomics) and other high-throughput technologies for studying molecular and genetic foundations of living systems' functioning has led to an information explosion in genetics, which is the main source of big data in world science, ahead of the other sciences and technologies in terms of the rate and volume of experimental data accumulation.

An important result of the analysis, interpretation and understanding of big genetic data is the formation of a new paradigm, wherein the main objects of genetics are not separate genes, but gene networks – groups of genes functioning in coordination, interacting with each other through their products, such as RNA, proteins, metabolites and other substances. It is gene networks that ensure the formation of all phenotypic (molecular, biochemical, cellular, physiological, morphological, behavioral, psychological, etc.) features of organisms based on the information coded in their genomes (Kolchanov et al., 2000, 2013; Ananko et al., 2002).

Reconstruction of gene networks is a very complex task requiring a search, extraction and integration of information scattered across tens of millions of scientific articles, thousands of factographic databases and millions of patents containing biological, medical, pharmacological, chemical, and other knowledge. To solve this task, it was necessary to develop computer software systems for automatic extraction of genetic data from the aforementioned sources using a combination of traditional textual analysis and methods of machine learning (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2022). To this day, more than 70,000 gene networks and their main components (signaling pathways, protein-protein, DNA-protein, RNA-protein interaction networks, metabolic pathways) have been reconstructed and presented in databases (Pico et al., 2008; Caspi et al., 2020; Kanehisa et al., 2023).

Accumulation of big data has resulted in the understanding of the great complexity of gene networks regulation on the base levels of their organization: each elementary fundamental biochemical or molecular biological process in a gene network is usually controlled by dozens, sometimes hundreds of elementary regulatory processes, whether it concerns protein enzyme activity, gene transcription regulation or "regulation of complex metabolic pathways" (Kolchanov et al., 2008). The abovementioned makes it incredibly difficult to reconstruct molecular mechanisms of the influence of genomic variability on phenotypic characteristics of organisms and clinical disease symptoms due to the fact that, among other things, regulatory processes are often characterized by a high degree of nonlinearity (Costanzo et al., 2019; Trifonova et al., 2021; Pratap et al., 2022) and dynamic instability in relation to changes in the initial data and constant physicochemical and molecular biological processes underlying the functioning of gene networks and regulatory systems (Khlebodarova et al., 2018).

Processing, analysis and interpretation of big genetic data streams requires the development of modern artificial intelligence methods focused on living systems. A key event that has initiated the rapid development of artificial intelligence methods in recent years is the creation of a new architecture of neural networks called transformers, which are geared towards the processing of symbol sequences, including texts in natural languages (Vaswani et al., 2017). The main feature of transformers is that the order of input sequences during processing is irrelevant. This provides ample opportunities for paralellizing, allowing for the deep learning of models on terabytes of data in a much shorter time than was previously possible using classic neural network architecture.

Let us note a few remarkable achievements of this approach. The creation of high-quality systems of machine translation from one natural language into another is of key importance (Jiao et al., 2023; Wang et al., 2023). The meaning of this development for science, technology, culture, art and human communication cannot be overestimated.

Based on transformer models, a huge breakthrough was made in solving one of the central tasks of molecular biology, which had been puzzling physicists, chemists and biologists for 60 years – predicting the spatial structure of globular proteins by their amino acid sequences. To solve this task, the AlphaFold (Thornton et al., 2021) and Rosetta (https://www.rosettacommons.org/) neural networks, predicting the 3D coordinates of heavy protein atoms with precision close to experimental, were developed. The network learning was based on hundreds of thousands of proteins with a known spatial structure and tens of millions amino acid sequences.

The methods of machine learning using transformer approaches created an opportunity for modeling the dynamics of complex molecular biological structures containing a large (up to 10^9) number of atoms (Pandey et al., 2022). These results are significant not only for fundamental science but also for a wide range of areas with a big potential for practical application, such as biotechnologies, genetics, medicine, pharmacology, creation of new materials, and many others.

Since 2017, when first publications on transformer technologies appeared, there has been an exponential growth of the number of publications using artificial intelligence methods (Eraslan et al., 2019; Boudry et al., 2022). Another machine learning approach that has been widely used and developed in the last years is graph neural networks (GNN), which provide entirely new opportunities for analysis of complex network structures based on the vector representation of graph vertices taking into account their local environment (Hamilton et al., 2017). The use of GNN is efficient for description, analysis and modeling of a wide range of network systems, be they natural, anthropogenic or technical: gene networks, intermolecular interaction networks, knowledge networks, social networks, etc. (Ektefaie et al., 2023).

In conclusion, it should be noted that there is a crucial limitation to a wide application of artificial intelligence methods in the areas of human activity that have a practical significance: its opaque decision-making process. In a number of works (Ma et al., 2018), a strategic way to overcome this restriction has been shown: it is necessary to develop hybrid information systems of a new generation, integrating classic methods of bioinformatics and systems computational biology and new artificial intelligence technologies based on the ontological description of the subject areas of research. In our opinion, only such an approach can ensure both the speed and quality of big genetic data processing with the use of artificial intelligence methods, and the transparency of the results obtained.

References

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. GeneNet: a database on structure and functional organisation of gene networks. *Nucleic Acids Res*. 2002;30(1):398-401. DOI 10.1093/nar/30.1.398

Boudry C., Al Hajj H., Arnould L., Mouriaux F. Analysis of international publication trends in artificial intelligence in ophthalmology. *Graefes Arch. Clin. Exp. Ophthalmol*. 2022;260(5):1779-1788. DOI 10.1007/s00417-021-05511-7

Caspi R., Billington R., Keseler I.M., Kothari A., Krummenacker M., Midford P.E., Ong W.K., Paley S., Subhraveti P., Karp P.D. The MetaCyc database of metabolic pathways and enzymes – a 2019 update. *Nucleic Acids Res*. 2020;48(D1):D445-D453. DOI 10.1093/nar/gkz862

Costanzo M., Kuzmin E., van Leeuwen J., Mair B., Moffat J., Boone C., Andrews B. Global genetic networks and the genotype-to-phenotype relationship. *Cell*. 2019;177(1):85-100. DOI 10.1016/j.cell.2019.01.033

Ektefaie Y., Dasoulas G., Noori A., Farhat M., Zitnik M. Multimodal learning with graphs. *Nat. Mach. Intell*. 2023;5:340-350. DOI 10.1038/s42256-023-00624-6

Eraslan G., Avsec Ž., Gagneur J., Theis F.J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet*. 2019;20(7):389-403. DOI 10.1038/s41576-019-0122-6

Hamilton W., Ying Z., Leskovec J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst*. 2017;30:1024-1034

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int. J. Mol. Sci*. 2022;23(23):14934. DOI 10.3390/ijms232314934

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. DOI 10.1186/s12859-018-2567-6

Jiao W., Wang W., Huang J.T., Wang X., Tu Z.P. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv*. 2023. DOI 10.48550/arXiv.2301.08745

Kanehisa M., Furumichi M., Sato Y., Kawashima M., Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 2023;51(D1):D587-D592. DOI 10.1093/nar/gkac963

Khlebodarova T.M., Kogai V.V., Trifonova E.A., Likhoshvai V.A. Dynamic landscape of the local translation at activated synapses. *Mol. Psychiatry*. 2018;23(1):107-114. DOI 10.1038/mp.2017.245

Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaia O.A., Ignatieva E.V., Goriachkovskaia T.N., Stepanenko E.L. Gene networks. *Molekulyarnaya Biologiya = Molecular Biology*. 2000;34(4):533-544 (in Russian)

Kolchanov N.A., Goncharov S.S., Likhoshvay V.A., Ivanisenko V.A. Systems Computational Biology. Novosibirsk: Publ. House SB RAS, 2008 (in Russian)

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvay V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vaviliv Journal of Genetics and Breeding*. 2013;4(2):833-850 (in Russian)

Ma J., Yu M.K., Fong S., Ono K., Sage E., Demchak B., Sharan R., Ideker T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*. 2018;15(4):290-298. DOI 10.1038/nmeth.4627

Pandey M., Fernandez M., Gentile F., Isayev O., Tropsha A., Stern A.C., Cherkasov A. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell*. 2022;4(3):211-221. DOI 10.1038/s42256-022-00463-x

Pico A.R., Kelder T., van Iersel M.P., Hanspers K., Conklin B.R., Evelo C. WikiPathways: pathway editing for the people. *PLoS Biol*. 2008;6(7):e184. DOI 10.1371/journal.pbio.0060184

Pratap A., Raja R., Agarwal R.P., Alzabut J., Niezabitowski M., Hincal E. Further results on asymptotic and finite-time stability analysis of fractional-order time-delayed genetic regulatory networks. *Neurocomputing*. 2022;475:26-37. DOI 10.1016/j.neucom.2021.11.088

Thornton J.M., Laskowski R.A., Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* 2021;27(10):1666-1669. DOI 10.1038/s41591-021-01533-0

Trifonova E.A., Klimenko A.I., Mustafin Z.S., Lashin S.A., Kochetov A.V. Do autism spectrum and autoimmune disorders share predisposition gene signature due to mTOR signaling pathway controlling expression? *Int. J. Mol. Sci.* 2021;22(10):5248. DOI 10.3390/ijms22105248

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *arXiv*. 2017. DOI 10.48550/arXiv.1706.03762

Wang L., Lyu C., Ji T., Zhang Z., Yu D., Shi S., Tu Z. Document-level machine translation with large language models. *arXiv*. 2023. DOI 10.48550/arXiv.2304.02210

Science editors of this issue:

N.A. Kolchanov, Full Member of the Russian Academy of Sciences, Academic Director of the Institute of Cytology and Genetics, SB RAS

Yu.G. Matushkin, Cand. Sc. (Biology), Lead Researcher of the Institute of Cytology and Genetics, SB RAS

COMPUTATIONAL GENOMICS
Original article

Вавиловский журнал генетики и селекции. 2023;27(7):728-736
DOI 10.18699/VJGB-23-85

Original Russian text https://vavilovj-icg.ru/

Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases

S.V. Filonov[1, 2], N.L. Podkolodnyy[1, 3] ✉, O.A. Podkolodnaya[1], N.N. Tverdokhleb[1], P.M. Ponomarenko[1], D.A. Rasskazov[1], A.G. Bogomolov[1], M.P. Ponomarenko[1]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pnl@bionet.nsc.ru

Abstract. It was previously shown that the expression levels of human genes positively correlate with TBP affinity for the promoters of these genes. In turn, single nucleotide polymorphisms (SNPs) in human gene promoters can affect TBP affinity for DNA and, as a consequence, gene expression. The Institute of Cytology and Genetics SB RAS (ICG) has developed a method for predicting TBP affinity for gene promoters based on a three-step binding mechanism: (1) TBP slides along DNA, (2) TBP stops at the binding site, and (3) the TBP-promoter complex is fixed due to DNA helix bending. The method showed a high correlation of theoretical predictions with measured values during repeated experimental testing by independent groups of researchers. This model served as a base for other ICG web services, SNP_TATA_Z-tester and SNP_TATA_Comparator, which make a statistical assessment of the SNP-induced change in the affinity of TBP binding to the human gene promoter and help predict changes in expression that may be associated with a genetic predisposition to diseases or phenotypic features of the organism. In this work, we integrated into a single database information about SNPs in human gene promoters obtained by automatic extraction from various heterogeneous data sources, as well as the estimates of TBP affinity for the promoter obtained using the three-step binding model and predicting their effect on gene expression for wild-type promoters and promoters with SNPs. We have shown that Human_SNP_TATAdb can be used for annotation and identification of candidate SNP markers of diseases. The results of a genome-wide data analysis are presented, including the distribution of genes with respect to the number of transcripts, the distribution of SNPs affecting TBP-DNA affinity with respect to positions within promoters, as well as patterns linking TBP affinity for the promoter, the specificity of the TBP binding site for the promoter and other characteristics of promoters. The results of the genome-wide analysis showed that the affinity of TBP for the promoter and the specificity of its binding site are statistically related to other characteristics of promoters important for the functional classification of promoters and the study of the features of differential gene expression.
Key words: TATA box; affinity; TBP; single nucleotide polymorphism; database; genome-wide analysis.

For citation: Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):728-736. DOI 10.18699/VJGB-23-85

Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования

С.В. Филонов[1, 2], Н.Л. Подколодный[1, 3] ✉, О.А. Подколодная[1], Н.Н. Твердохлеб[1], П.М. Пономаренко[1], Д.А. Рассказов[1], А.Г. Богомолов[1], М.П. Пономаренко[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ pnl@bionet.nsc.ru

Аннотация. Ранее было показано, что уровень экспрессии генов человека положительно коррелирует с аффинностью TBP к промоторам этих генов. В свою очередь, однонуклеотидные полиморфизмы (SNP) в промоторах

генов человека могут влиять на аффинность белка TBP к ДНК и, как следствие, на экспрессию генов. В ИЦиГ СО РАН разработан метод предсказания аффинности TBP к промоторам генов на основе трехшагового механизма связывания, включающего скольжение TBP по ДНК, остановку TBP в месте связывания, фиксацию комплекса TBP–промотор за счет изгиба спирали ДНК. Метод показал высокую корреляцию теоретических предсказаний с измеренными значениями при многократной экспериментальной проверке независимыми группами исследователей. На основе этой модели в ИЦиГ СО РАН ранее были разработаны веб-сервисы SNP_TATA_Z-tester и SNP_TATA_Comparator, позволяющие вычислять статистическую оценку вызванного SNP изменения аффинности связывания TBP с промотором гена человека и прогнозировать изменение экспрессии, которые могут быть связаны с генетической предрасположенностью к заболеваниям или фенотипическими особенностями организма. В настоящей работе проведена интеграция в единой базе данных информации об однонуклеотидных полиморфизмах в промоторах генов человека, полученной путем автоматической экстракции из различных гетерогенных источников данных, а также результатов оценки аффинности TBP к промотору с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом. Показана возможность использования базы данных Human_SNP_TATAdb для аннотации и выявления кандидатных SNP-маркеров заболеваний. Представлены результаты полногеномного анализа данных, включая особенности распределения генов по количеству транскриптов, распределение SNP, влияющих на аффинность TBP к ДНК по позициям внутри промоторов, а также закономерности, связывающие между собой аффинность TBP к промотору, специфичность сайта связывания TBP с промотором и другие характеристики промоторов. Результаты полногеномного анализа показали, что аффинность TBP к промотору и специфичность его сайта связывания статистически связаны с другими характеристиками промоторов, важными для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов.

Ключевые слова: TATA-бокс; аффинность; TBP; однонуклеотидный полиморфизм; база данных; полногеномный анализ.

Introduction

The development of methods for predicting the effect of mutations on the level of gene expression for various organisms is important for solving many problems in the field of biotechnology, plant breeding, medicine, etc. Mutations in the human genome can be associated with a variety of physiological characteristics and diseases, and knowledge of their presence and cause is certainly necessary for the actively developing approach of personalized medicine.

The most common type of mutation in the human genome is SNP (Single Nucleotide Polymorphism), which is a single nucleotide difference in the DNA sequence. SNPs can be localized in different functional regions of the genome, which determines the nature of their influence. Mutations in the coding regions of the gene are the most studied; they directly affect the structure of the transcribed mRNA and, consequently, the synthesized protein. However, genome-wide association studies (GWAS) have shown that most SNPs that are significantly associated with disease susceptibility are located in non-coding regions (Hindorff et al., 2009; French and Edwards, 2020; Chandra et al., 2021), and more than 90 % of them are located in regulatory elements (Maurano et al., 2012). At the moment, one of the most studied regulatory regions is the TATA box region in the promoter, the sequence of which determines the affinity of the TBP protein (TATA binding protein), which is a key transcription initiation factor. Mutations in this region can affect the binding of the TBP protein to the promoter and, consequently, gene expression (Savinkova et al., 2007).

Previously, at the Institute of Cytology and Genetics SB RAS, a method for predicting the affinity of TBP for gene promoters based on a three-step binding mechanism was developed (Ponomarenko et al., 2008). The method showed a high correlation of theoretical predictions with experimentally measured affinity values when tested multiple times by independent groups of researchers (Delgadillo et al., 2009; Savinkova et al., 2013; Oshchepkov et al., 2022). Based on this model, the Institute of Cytology and Genetics SB RAS developed the SNP_TATA_Z-tester web service (Rasskazov et al., 2013), which allows one to calculate a statistical assessment of the SNP-induced change in the binding affinity of TBP for the human gene promoter and predict changes in expression. Using this web service, we previously identified candidate SNP markers for autoimmune diseases (Ponomarenko et al., 2016a), behavioral disorders (Chadaeva et al., 2016), chronopathologies (Ponomarenko et al., 2016b) and other diseases.

In this work, we integrated into a database information about SNPs in human gene promoters, obtained by automatic extraction from various heterogeneous data sources, as well as the results of assessing the affinity of TBP for the promoter and the specificity of the TBP binding site using a three-step binding model and assessing their effect on gene expression for the reference genome promoters and promoters with SNPs.

The main use of the Human_SNP_TATAdb database is the annotation of promoters and genes in order to identify candidate SNP markers of diseases. Considering that quite a lot of research that includes this kind of annotation has already been carried out, we present one of the options as an example.

This article presents the results of a genome-wide data analysis, including features of the distribution of genes by the number of transcripts, the distribution of SNPs affecting the affinity of TBP for DNA by positions within promoters. The article also presents patterns connecting the affinity of TBP

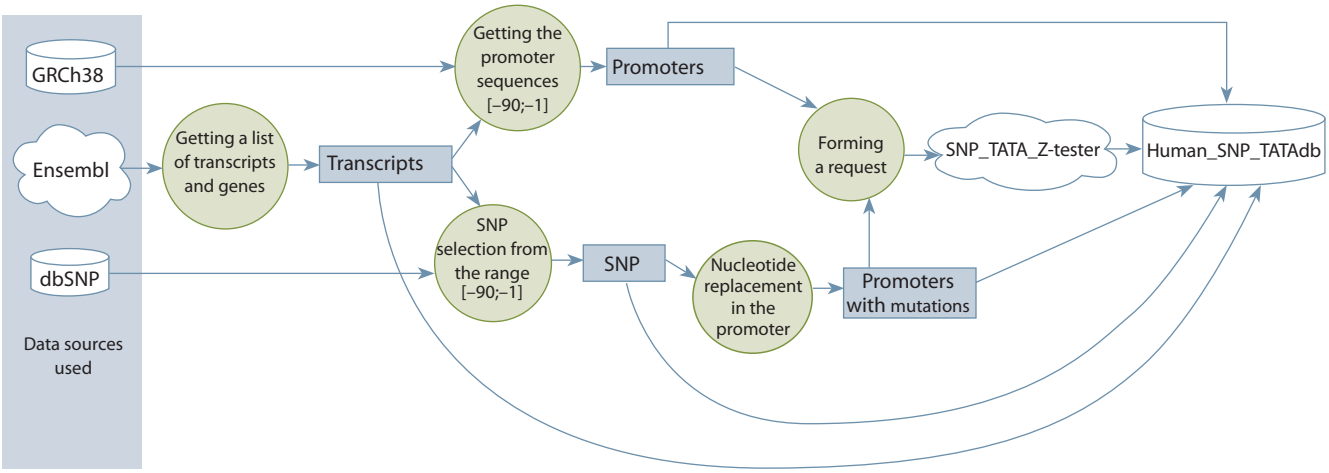


Fig. 1. Data flow diagram for initializing the Human_SNP_TATAdb database.

for the promoter, the specificity of the binding site of TBP for the promoter, and other characteristics of promoters that are important for the functional classification of promoters and the study of features of differential gene expression.

Materials and methods

Below, we present a data flow diagram for data integration and database initialization (Fig. 1). Further, all stages of work are described in more detail. Data on genes and their attributes, transcription starts and transcripts were obtained from the Ensembl web service (Birney et al., 2004). To access the services and databases used in the work, the Bioconductor library of the R language was used, with the following packages:

1. biomaRt[1] is a package that provides an interface to the ENSEMBL collection of databases, allowing large volumes of data to be retrieved in a unified way and used in data analysis in Bioconductor.
2. BSgenome.Hsapiens.NCBI.GRCh38[2] is a package that provides access to the Homo sapiens (Human) genome sequence provided by NCBI (GRCh38.p13).
3. SNPlocs.Hsapiens.dbSNP155.GRCh38[3] is a dbSNP 155 access package including information on 949,021,448 SNPs in chromosomes 1–22, X, Y and MT that was extracted from RefSNP JSON files.

To identify the start of transcription, it is necessary to use transcripts with high-quality annotation that includes this information and for which there is evidence of their biological relevance. When annotating transcripts in Ensembl, special tags identify the highest quality annotated transcripts. We included in the database only those transcripts, the annotation quality of which corresponds to the "GENCODE Basic" label[4]. According to the specification, the Enseml GENCODE Basic set contains at least one transcript for each gene in the GENCODE genetic set, regardless of biotype, i.e. each gene is represented in the core GENCODE set. For protein-coding genes, only full-length protein-coding transcripts are included in the core GENCODE set.

For the specified transcription start coordinates, the coordinates and nucleotide sequences of the corresponding promoter are determined ([–90; –1] from the transcription start). We obtained SNP data using the dbSNP database[5] (Sherry et al., 2001). For each promoter, SNPs located within [–90; –1] from the start of transcription were identified. Minor promoter sequence variants were created automatically by adding corresponding nucleotide substitutions from the dbSNP database (issue 155) to the major sequence variants.

The Bucher weight matrix (Bucher, 1990) was used to identify TATA-containing promoters.

The affinity of TBP for DNA was calculated using a three-step binding model previously developed at the Institute of Cytology and Genetics SB RAS (Ponomarenko et al., 2008) and a multi-threaded high-performance version of the SNP_TATA_Z-tester program also implemented by us. This program also allows you to evaluate the statistical significance of changes in the affinity of the TBP protein for the promoter due to point nucleotide substitutions (SNPs) in the promoter using a z-test.

The affinity of TBP is described by the association constant of the TBP/DNA complex. However, at present, instead of the association constant, the inverse measure is usually used – the dissociation constant K_d. In this case, the affinity of TBP for DNA, measured in nanomoles per liter (nM/L), will be equal to $A = 10^9/K_d$. The lower the Kd, the higher the affinity of TBP for the promoter and the stronger the interaction of TBP with the promoter.

The second option presented in the database is the logarithmic form of affinity $\alpha = 9*\ln(10) - \ln(K_d)$, which is convenient for comparing TBP affinity to the promoter, since it has a distribution close to normal. As α increases, the affinity of TBP for the promoter and the strength of their interaction increase.

[1] https://bioconductor.org/packages/release/bioc/html/biomaRt.html
[2] https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.NCBI.GRCh38.html
[3] https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP155.GRCh38.html
[4] https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html
[5] https://www.ncbi.nlm.nih.gov/snp/

We performed affinity calculations for reference DNA sequences of all promoters and minor sequence variants of these promoters with one single nucleotide polymorphism. For each minor sequence, we assessed the deviation of TBP affinity for the promoter from the affinity obtained for the promoter DNA sequence from the reference genome. At the same time, the level of statistical significance of these changes was determined.

It was previously shown that the affinity of TBP for the promoter is statistically significantly correlated with the level of expression of the corresponding transcript (Mogno et al., 2010). Therefore, with a statistically significant increase or decrease in TBP affinity, an estimate of the corresponding change in the level of transcript expression is indicated in the database.

Based on the estimates of TBP protein-promoter affinity, we introduced additional characteristics, such as TBP protein-promoter binding site specificity, which are useful for promoter classification and biological annotation of groups of promoters or genes.

The specificity of the binding site of TBP for the gene promoter corresponds to the maximum normalized affinity of TBP for the gene promoter relative to the average affinity of TBP for each position of the sliding window (Ponomarenko et al., 2015), not including 10 positions closest to the start of transcription (55 values in total). Specificity Z was calculated as follows:

$$Z = \frac{\alpha_{\max} - \overline{\alpha}}{\sigma_\alpha}, \quad \sigma_\alpha = \sqrt{\frac{1}{54}\sum_1^{55}(\alpha_i - \overline{\alpha})},$$

where α_i is the assessment of the affinity of TBP to the promoter at position i, $\overline{\alpha}$ is the average value of α_i, σ_α is the unbiased estimate of standard deviation α_i, Z is the specificity of the TBP protein binding site for the promoter.

Another important indicator describing the change in the affinity of TBP for the promoter caused by SNP is the natural logarithm of the K_d ratio for the reference (*wt*) and minor (*mt*) alleles of the SNP in question, which is calculated as follows:

$$k_{snp} = \ln(K_{d,\,wt}/K_{d,\,mt}).$$

Positive or negative k_{snp} values indicate that gene expression for the minor allele is, respectively, higher or lower than for the reference variant. This score was used to identify candidate SNP markers that may be associated with genetic susceptibility to diseases; in particular, we made predictions that are consistent with clinical evidence of underexpression of this gene in patients with variable immunodeficiency, stroke and preeclampsia (Ponomarenko et al., 2017).

Results and discussion

Database

The Human_SNP_TATAdb database has been developed, its logical diagram is presented in Fig. 2. The database was populated in accordance with the data integration and database initialization scenario presented in Fig. 1.

The database is implemented on the basis of the MySQL DBMS[6] version 8.0 and includes 6 main tables (chromosomes, genes, transcripts, snps, promoters, promoters_has_snps), 10 uxiliary tables and dictionaries. The diagram of the developed database is shown in Fig. 2. Queries to the database are carried out through SQL queries.

The chromosomes table includes the chromosome identifier, length, number of nucleotides, and species of organism.

The genes table includes information about gene identifiers in various databases, including ensembl, gene symbol name, chromosome reference, chain, and gene biotype.

The transcripts table includes information about transcript identifiers, transcript coordinates in the genome, transcript biotype, and a link to the promoter and the gene.

The snps table includes the following information: SNP identifiers, SNP positions in the genome, chromosome reference and allele. Here and below, one SNP is taken to be an unambiguous variant of a genome change. Polymorphisms that have one rs identifier, but allow several nucleotide substitution options, are counted by the number of such options.

It should be noted that the same nucleotide substitution can occur in different gene promoters and differently change the level of affinity of the TBP protein for these promoters, and therefore two tables are defined in the database to describe the promoters, promoters and promoters_has_snps, with a 1:N ratio (one promoter can influence several SNPs), and the snps and promoters_has_snps tables are also related by a 1:N relationship (one SNP can be included in several promoters).

The first promoters table includes the following information: promoter identifier, DNA sequence corresponding to the region [–90; –1] from the start of transcription, coordinates of the start and end of the promoter in the genome, affinity of the TBP protein for the promoter with an error, link to the gene.

The promoters_has_snps table includes information about the promoter identifier, a link to SNP, coordinates of SNP in the promoter and relative to the start of transcription, the sequence of the wild-type promoter and the promoter with SNP, the affinity of TBP for the promoter with an error, the nature of changes in gene expression due to mutation in the promoter, the significance level of the statistical test.

The source_snp_dbs table includes information about data sources, database versions, links to databases, which is necessary for automated updating of the Human_SNP_TATAdb database.

Table relationship types define constraints that match the provenance of the data and are therefore important for maintaining the integrity of the database, as well as for providing additional control over the data and reducing the possibility of errors. In particular, each gene may have one or more promoters, and each promoter may regulate the expression of one or more transcripts.

As a result, the database contains the following information:

- 62603 genes, of which 19314 encode proteins.
- 117414 transcripts, of which 63141 encode proteins.
- 5,305,816 SNP variants in gene promoters in the [–90; –1] interval from the start of transcription, of which 3,199,285 are in the promoters of protein-coding genes.
- For 445,875 SNP variants in the promoter of a protein-coding gene, we predicted that they statistically significantly (p-value < 0.05) change the level of TBP affinity for this promoter.

[6] https://www.mysql.com/

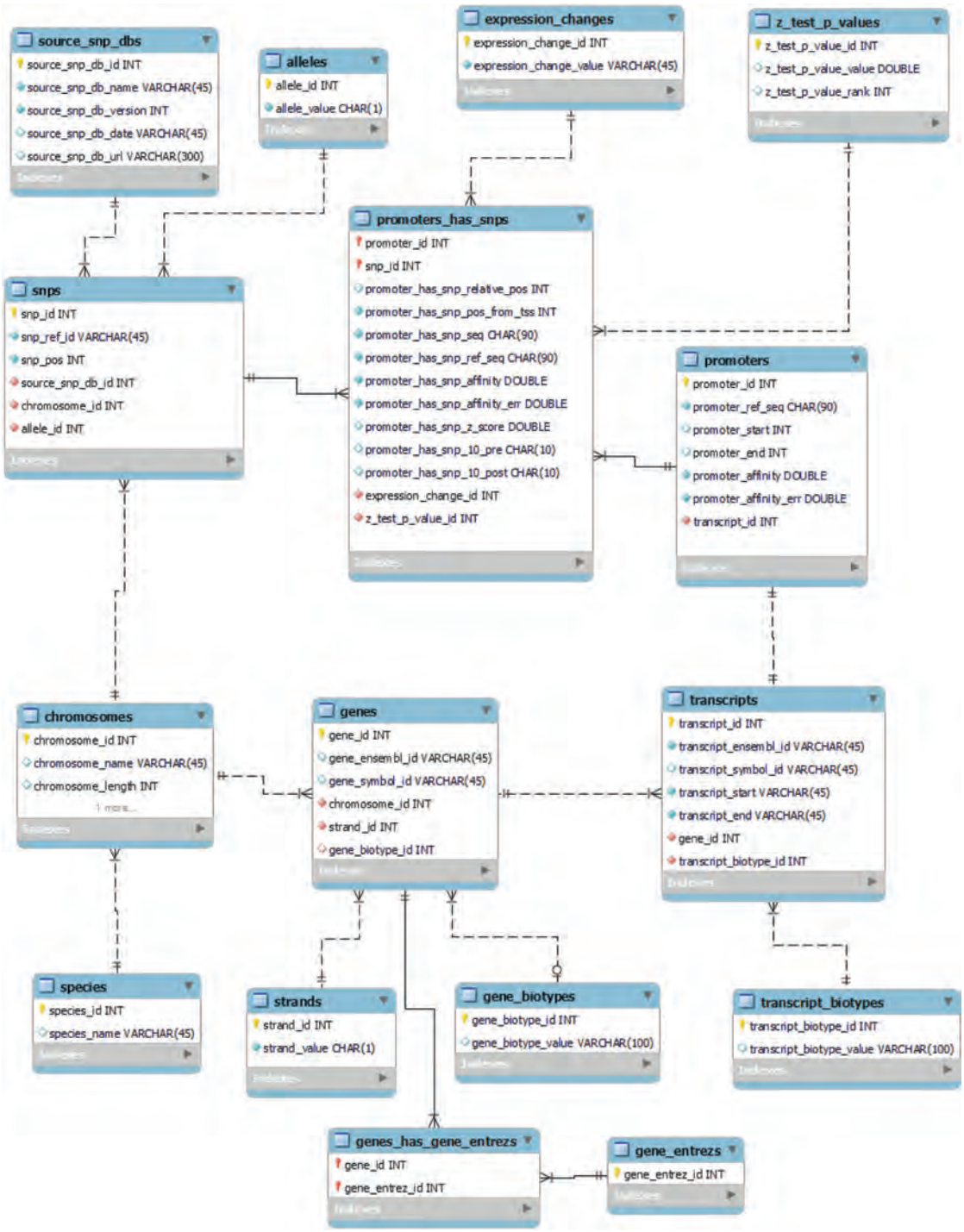


Fig. 2. Scheme of the Human_SNP_TATAdb database.

Application options for the Human_SNP_TATAdb database

The information presented in the database (affinity of the TBP protein for the promoter, specificity of the binding site of TBP for the promoter and assessment of changes in these characteristics due to SNP) may be important for identification of markers of genetic susceptibility to diseases, identification and functional interpretation of classes of promoters similar in the mechanism of regulation of the early stage transcription initiation, etc.

The Human_SNP_TATAdb database can also help to annotate genes or a group of genes in terms of TBP affinity for a promoter or TBP binding site specificity for a promoter. To determine the characteristic of a gene associated with the specific binding of TBP to gene promoters for the purpose of GO analysis, you can use the average values of the affinity of TBP for gene promoters or the affinity of TBP for the promoter

corresponding to the only transcript for the gene, which is determined by ENSEMBL experts as canonical and is specified in the database with the label "Ensembl Canonical"[7], i.e. it is generally the most conserved, the most expressed, has the longest coding sequence, and is represented in other key resources such as NCBI and UniProt. We mark its corresponding promoter as canonical and use characteristics such as the affinity of TBP for the canonical promoter and the specificity of the TBP binding site for the canonical promoter to annotate a gene or group of genes.

Correlation analysis showed that there is a strong linear relationship between the affinity of TBP for the canonical gene promoter and the average affinity of gene promoters ($R = 0.88$, d.f. = 19308). Therefore, using any option will lead to similar results. However, using TBP's affinity for the canonical gene promoter appears to be biologically more reasonable. Of course, a key use case for the Human_SNP_TATAdb database is gene annotation and identification of candidate SNP markers for disease susceptibility.

Considering that to date quite a lot of studies have already been conducted in which this kind of annotation has been carried out, we will present the work (Bogomolov et al., 2023) as an example using the Human_SNP_TATAdb database for annotation and identification of candidate SNP markers of atherogenesis, atherosclerosis and atheroprotection.

We pre-selected 1068 human genes associated with these diseases. Information about single nucleotide polymorphisms in the promoters of these human genes, the results of assessing the affinity of TBP for promoters and assessing their effect on gene expression for wild-type promoters and promoters with SNP was obtained from the Human_SNP_TATAdb database. This information was supplemented by an annotation of selected genes prepared by experts, and a database view was generated, focused on the analysis of genes associated with atherogenesis, atherosclerosis and atheroprotection, external access to which is provided via the Web interface[8].

In silico analysis of all 5112 SNPs in their promoters identified 330 candidate SNP markers that statistically significantly alter the affinity of TBP for these promoters.

Next, we compared the corresponding frequencies of SNPs that increase and decrease the affinity of TBP for the promoters of the same genes. This comparison was made to analyze whether these genes are under the influence of natural selection or neutral drift. We found that natural selection acts against underexpression of hub genes for atherogenesis, atherosclerosis and atheroprotection and, through enhanced atheroprotection, contributes to improved human health (Bogomolov et al., 2023).

Examples of application of the Human_SNP_TATAdb database for genome-wide analysis

The developed database makes it possible to analyze genome-wide statistics and the distribution of these indicators in various groups of promoters, for example, TATA-containing promoters. For genome-wide analysis, we used protein-coding genes and transcripts selected by the values of the 'gene_biotype' and 'transcript_biotype' fields equal to 'protein_coding'.

[7] https://www.ensembl.org/info/genome/genebuild/canonical.html
[8] http://www.sysbio.ru/Human_SNP_TATAdb

Alternative promoters and TBP/DNA affinity

It should be noted that one gene can have several transcripts, the initiation of transcription of which occurs using different promoters, for which the affinity of the TBP protein is assessed. Figure 3 shows the distribution of protein-coding genes by transcript number. The largest number of protein-coding genes (29.77% of genes) have a single transcript and, as a consequence, one promoter. 5% of protein-coding genes have at least 9 protein-coding transcripts. Analysis of the distribution of genes by the number of transcripts showed that the average number of transcripts per gene is 3.27, and the median is 2 transcripts per gene. The *Mapk10 (mitogen-activated protein kinase 10)* gene has the maximum number of protein-coding transcripts (87).

Our analysis showed that the distribution of the average affinity of TBP for canonical promoters in groups of genes divided by the number of transcripts is close to uniform. Thus, there is no need to neutralize the effects due to different numbers of transcripts per gene when analyzing data using TBP affinity.

Distribution of SNPs that change gene expression by promoter positions

The distribution of SNPs that statistically significantly change gene expression at positions from the start of transcription is clearly different from uniform (Fig. 4). In the region [–35; –20], corresponding to the usual location of the TATA box, the number of such SNPs is noticeably higher than in other regions of the promoter. The number of SNPs that reduce gene expression in the [–35; –20] region, corresponding to the location of the TATA-box, is more than one and a half higher than in other regions of the promoter. This may be due to the fact that SNPs in this region tend to disrupt the TATA box.

The number of SNPs that increase gene expression is higher on the flanks of the most frequent TATA box locations. The peaks are located at positions –24 and –32 from the start of transcription. It should be noted that the distribution of all SNPs across the promoter positions of protein-coding genes is uniform. This suggests that an increase in the number of SNPs that increase gene expression on the flanks of the TATA box may have functional significance.

Affinity of TBP to TATA-containing and TATA-free promoters of protein-coding genes

Analysis of the dependence of TBP/DNA affinity indicators, measured on a logarithmic scale ($\alpha = 9*\ln(10) - -\ln(K_d)$, for TATA-containing and TATA-free promoters of protein-coding genes (Fig. 5), showed that the group of TATA-containing promoters exhibits higher TBP/DNA affinity, consistent with stronger TBP-promoter affinity.

Functional SNPs affecting the affinity of TBP for promoter DNA and the specificity of the TBP protein binding site

We analyzed the dependence of the proportion of SNPs that have a statistically significant effect on the affinity of TBP for the DNA of the promoters of protein-coding genes on the specificity of the TBP protein binding site (Fig. 6). It has been shown that SNPs in promoters with low specificity of the

Fig. 3. Distribution of protein-coding genes by number of transcripts.

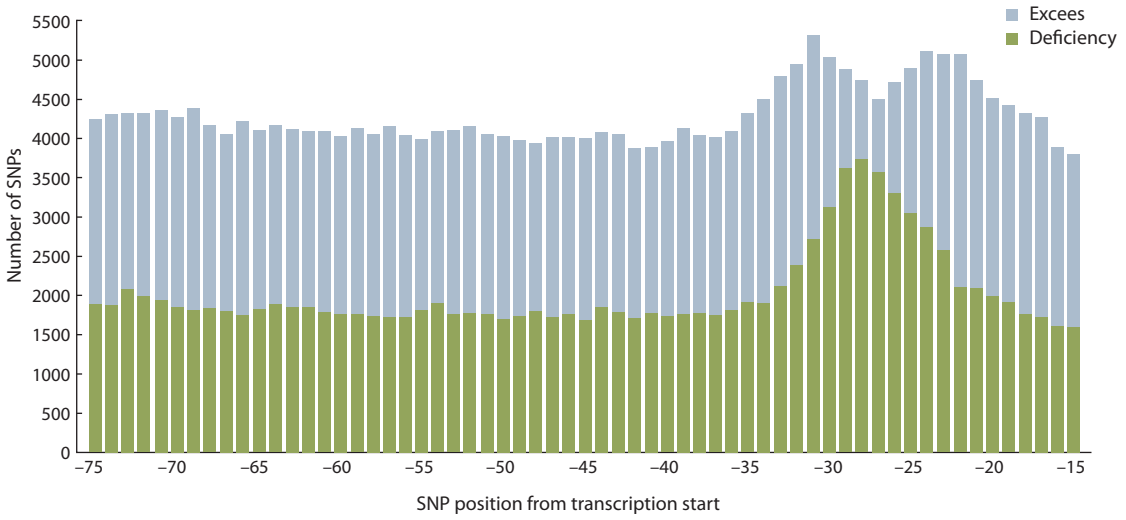


Fig. 4. Distribution of the number of SNPs that increase (excess) and decrease (deficiency) the affinity of TBP for the DNA of the promoters of protein-coding genes, depending on the position of the SNP relative to the start of transcription.

TBP binding site for the promoter, as a rule, lead to an increase in gene expression, and in promoters with high specificity, the proportion of SNPs that decrease expression is increased.

Analysis of the contingency table showed that low specificity values of the TBP binding site to the promoter (spec less than 2.5) are more often observed in promoters without a TATA box (TATA–) ($\chi^2 = 10385$, p-value < 1.0e–228).

Conclusion

This work presents the Human_SNP_TATAdb database, which includes information on single nucleotide polymorphisms in human gene promoters obtained by automatic extraction from various heterogeneous data sources, the results of assessing the affinity of TBP for the promoter using a three-step binding model, and assessing their impact on gene expression for wild-type promoters and promoters with a single nucleotide polymorphism.

The affinity of the TBP protein for the promoter, the specificity of the TBP binding site for the promoter, and assessments of changes in these characteristics with single nucleotide polymorphisms presented in the database may be important for identification of candidate markers of genetic susceptibility to diseases, identification and functional interpretation of classes of promoters that are similar in the mechanism of regulation of the early stage of transcription initiation, etc.

The Human_SNP_TATAdb database can also be used to annotate genes or groups of genes in terms of TBP-promoter affinity or TBP-promoter binding site specificity.

The results of genome-wide analysis showed that the affinity of TBP for the promoter and the specificity of its binding site are statistically associated with other characteristics of promoters that are important for the functional classification of promoters and the study of differential gene expression patterns.

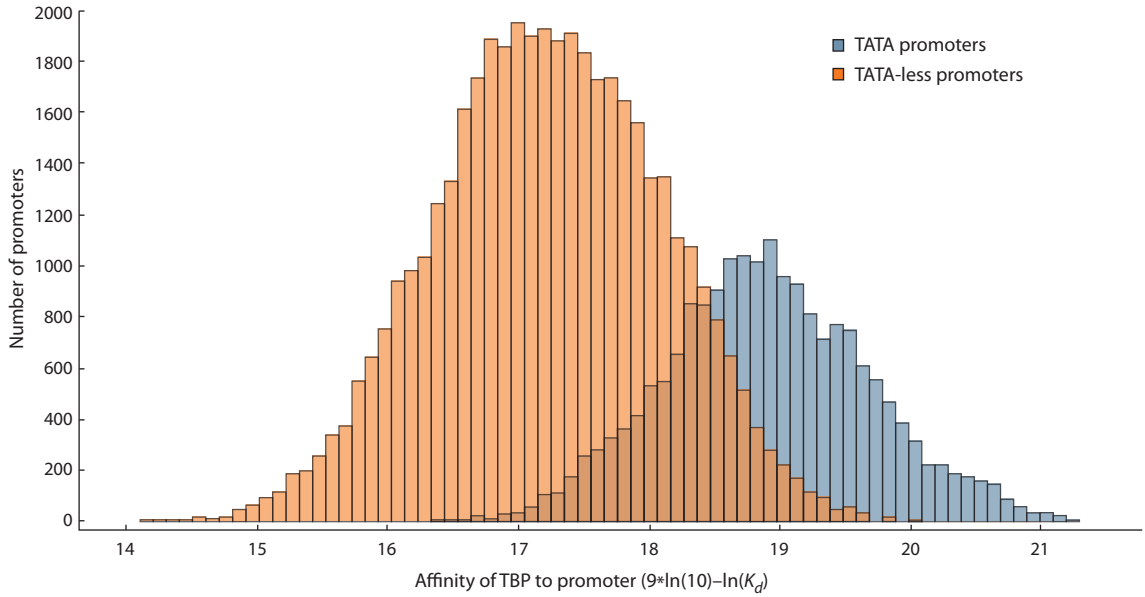


Fig. 5. Distribution of promoters of protein-coding genes by TBP affinity in groups of TATA-containing promoters and promoters without a TATA box. The x-axis TBP affinity score for the promoter is given on a logarithmic scale.

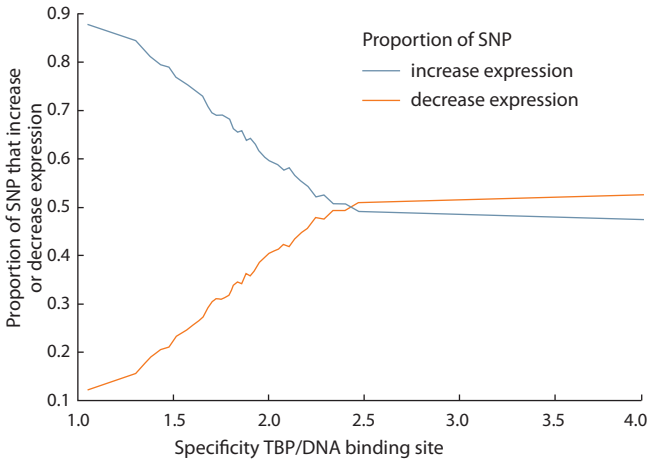


Fig. 6. Proportion of SNPs in promoters that increase and decrease the expression of protein-coding genes depending on the specificity of the TBP binding site for the DNA promoter.

Contingency table of the specificity of the TBP binding site with the promoter and the presence of a TATA box in the promoter

Specificity	TATA–	TATA+	Total
Spec < 2.5	**29114**	10379	39493
Spec ≥ 2.5	14538	9109	23647
Total	43652	19488	63140

The use of the Human_SNP_TATAdb database for gene annotation and the identification of candidate SNP markers of atherogenesis, atherosclerosis and atheroprotection is one example, as a result of which new knowledge is emerging about the effect of various single polymorphisms on susceptibility to certain diseases.

References

Birney E., Andrews T.D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., ..., Cox A., Hubbard T., Clamp M. An overview of Ensembl. *Genome Res*. 2004;14(5):925-928. DOI 10.1101/gr.1860604

Bogomolov A., Filonov S., Chadaeva I., Rasskazov D., Khandaev B., Zolotareva K., Kazachek A., … Kolchanov N., Tverdokhleb N., Ponomarenko M. Candidate SNP markers significantly altering the affinity of TATA-binding protein for the promoters of human hub genes for atherogenesis, atherosclerosis and atheroprotection. *Int. J. Mol. Sci*. 2023;24(10):9010. DOI 10.3390/ijms24109010

Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*. 1990;212(4):563-578. DOI 10.1016/0022-2836(90)90223-9

Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A., Sharypova E.B., Kashina E.V., Matveeva M.Yu., Arshinova T.V., Ponomarenko P.M., Arkova O.V., Bondar N.P., Savinkova L.K., Kolchanov N.A. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics*. 2016;17(Suppl. 14):995. DOI 10.1186/s12864-016-3353-3

Chandra V., Bhattacharyya S., Schmiedel B.J., Madrigal A., Gonzalez-Colin C., Fotsing S., Crinklaw A., Seumois G., Mohammadi P., Kronenberg M., Peters B., Ay F., Vijayanand P. Promoter interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat. Genet*. 2021;53(1):110-119. DOI 10.1038/s41588-020-00745-3

Delgadillo R.F., Whittington J.E., Parkhurst L.K., Parkhurst L.J. The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. *Biochemistry*. 2009; 48(8):1801-1809. DOI 10.1021/bi8018724

French J.D., Edwards S.L. The role of noncoding variants in heritable disease. *Trends Genet.* 2020;36(11):880-891. DOI 10.1016/j.tig.2020.07.004

Hindorff L.A., Sethupathy P., Junkins H.A., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*. 2009;106(23):9362-9367. DOI 10.1073/pnas.0903103106

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., … Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. DOI 10.1126/science.1222794

Mogno I., Vallania F., Mitra R.D., Cohen B.A. TATA is a modular component of synthetic promoters. *Genome Res*. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Oshchepkov D., Chadaeva I., Kozhemyakina R., Zolotareva K., Khandaev B., Sharypova E., Ponomarenko P., Bogomolov A., Klimova N.V., Shikhevich S., Redina O., Kolosova N.G., Nazarenko M., Kolchanov N.A., Markel A., Ponomarenko M. Stress reactivity, susceptibility to hypertension, and differential expression of genes in hypertensive compared to normotensive patients. *Int. J. Mol. Sci*. 2022;23(5):2835. DOI 10.3390/ijms23052835

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl. Biochem. Biophys*. 2008;419:88-92. DOI 10.1134/S1607672908020117

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed Res. Int*. 2015;2015:359835. DOI 10.1155/2015/359835

Ponomarenko M.P., Arkova O., Rasskazov D., Ponomarenko P., Savinkova L., Kolchanov N. Candidate SNP markers of genderbiased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front. Immunol*. 2016a;7:130. DOI 10.3389/fimmu.2016.00130

Ponomarenko P., Rasskazov D., Suslov V., Sharypova E., Savinkova L., Podkolodnaya O., Podkolodny N.L., Tverdokhleb N.N., Chadaeva I., Ponomarenko M., Kolchanov N. Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed Res. Int*. 2016б;2016:8642703. DOI 10.1155/2016/8642703

Ponomarenko M., Rasskazov D., Chadaeva I., Sharypova E., Ponomarenko P., Arkova O., Kashina E., Ivanisenko N., Zhechev D., Savinkova L., Kolchanov N. SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* (*Schol. Ed.*). 2017;9(2):276-306. DOI 10.2741/s488

Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_COMPARATOR: web service for comparison of SNPS within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/1):599-606 (in Russian)

Savinkova L.K., Drachkova I.A., Ponomarenko M.P., Lysova M.V., Arshinova T.V., Kolchanov N.A. Interaction of recombinant TATA-binding protein with mammals gene promoter TATA boxes. *Ekologicheskaya genetika = Ecological genetics*. 2007;5(2):44-49. DOI 10.17816/ecogen5244-49 (in Russian)

Savinkova L., Drachkova I., Arshinova T., Ponomarenko P., Ponomarenko M., Kolchanov N. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PLoS One*. 2013;8(2).e54626. DOI 10.1371/journal.pone.0054626

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311. DOI 10.1093/nar/29.1.308

ORCID ID

N.L. Podkolodnyy orcid.org/0000-0001-9132-7997
O.A. Podkolodnaya orcid.org/0000-0003-3247-0114
P.M. Ponomarenko orcid.org/0000-0003-2715-9612
D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
M.P. Ponomarenko orcid.org/0000-0003-1663-318X

Acknowledgements. The work was supported by budget projects FWNR-2022-0020, No. 0251-2022-0005 and the Federal Scientific and Technical Program for the Development of Genetic Technologies of Russia

Conflict of interest. The authors declare no conflict of interest.

Received August 22, 2023. Revised September 15, 2023. Accepted September 19, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):737-745
DOI 10.18699/VJGB-23-86

Original Russian text https://vavilovj-icg.ru/

GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing

A.Y. Pronozin[1, 2] ✉, E.A. Salina[1, 2, 3], D.A. Afonnikov[1, 2, 4]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State Agrarian University, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
✉ pronozinartem95@gmail.com

Abstract. The development of next-generation sequencing technologies has provided new opportunities for genotyping various organisms, including plants. Genotyping by sequencing (GBS) is used to identify genetic variability more rapidly, and is more cost-effective than whole-genome sequencing. GBS has demonstrated its reliability and flexibility for a number of plant species and populations. It has been applied to genetic mapping, molecular marker discovery, genomic selection, genetic diversity studies, variety identification, conservation biology and evolutionary studies. However, reduction in sequencing time and cost has led to the need to develop efficient bioinformatics analyses for an ever-expanding amount of sequenced data. Bioinformatics pipelines for GBS data analysis serve the purpose. Due to the similarity of data processing steps, existing pipelines are mainly characterised by a combination of software packages specifically selected either to process data for certain organisms or to process data from any organisms. However, despite the usage of efficient software packages, these pipelines have some disadvantages. For example, there is a lack of process automation (in some pipelines, each step must be started manually), which significantly reduces the performance of the analysis. In the majority of pipelines, there is no possibility of automatic installation of all necessary software packages; for most of them, it is also impossible to switch off unnecessary or completed steps. In the present work, we have developed a GBS-DP bioinformatics pipeline for GBS data analysis. The pipeline can be applied for various species. The pipeline is implemented using the Snakemake workflow engine. This implementation allows fully automating the process of calculation and installation of the necessary software packages. Our pipeline is able to perform analysis of large datasets (more than 400 samples).
Key words: genotyping by sequencing (GBS); bioinformatic pipeline; hordeum.

For citation: Pronozin A.Y., Salina E.A., Afonnikov D.A. GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023; 27(7):737-745. DOI 10.18699/VJGB-23-86

GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования

А.Ю. Пронозин[1, 2] ✉, Е.А. Салина[1, 2, 3], Д.А. Афонников[1, 2, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский государственный аграрный университет, Новосибирск, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ pronozinartem95@gmail.com

Аннотация. Развитие технологий секвенирования нового поколения открыло новые возможности для генотипирования различных организмов, включая растения. Метод генотипирования путем секвенирования (GBS) применяется для идентификации генетической изменчивости и более быстрого генотипирования образцов, а также является более экономически эффективным методом в сравнении с полногеномным секвенированием. GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Этот метод был применен для генетического картирования, выявления молекулярных маркеров, геномной селекции, в исследовании генетического разнообразия, идентификации сортов, а также в исследованиях в области биологии охраны природы и эволюционной экологии. Однако сокращение времени и стоимости секвенирования привело к необходимости разработки качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных. Для этих целей были разработаны биоинформатические конвейеры анализа данных, полученных методом GBS. Вследствие схожести этапов обработки существующие конвейеры

в основном различаются комбинацией программных пакетов, специфически подобранных для обработки данных как для определенных, так и для любых организмов. Несмотря на качественно подобранные пакеты программ, конвейеры имеют некоторые недостатки, например отсутствие возможности автоматизации процесса расчета (каждый этап нужно запускать вручную), что значительно снижает скорость исследования. В большинстве конвейеров отсутствует возможность автоматической установки всех необходимых программных пакетов, а также нет возможности отключения ненужного или пройденного этапа. В настоящей работе нами был разработан биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер применим для любых видов организмов. Реализация конвейера на платформе Snakemake позволила полностью автоматизировать процесс расчета и установки необходимых программных пакетов. Конвейер позволяет обрабатывать большие объемы данных (более 400 образцов).
Ключевые слова: генотипирование путем секвенирования; биоинформатический конвейер; ячмень.

Introduction

Genetic diversity is the most important basis for studying plant resistance to biotic and abiotic stresses and for developing new highly adaptive and high-yielding crop sorts. Study of genetic diversity is performed using various methods of DNA analysis. To date, one of the most advanced methods is the use of molecular markers (Kanukova et al., 2019). Molecular markers (DNA markers) are genetic markers analysed at the nucleotide level (Khlestkina, 2013). Their use allows to identify genetic diversity of populations, subspecies, species, allowing to effectively determine loci controlling economically valuable plant traits even at the initial stage of breeding (Sukhareva, Kuluev, 2018).

Some of the most convenient DNA markers for genetic analysis are SNP markers (Khlestkina, 2013). SNP (Single-Nucleotide Polymorphism) is a single-nucleotide position in genomic DNA for which different sequence variations (alleles) occur in the population (Sukhareva, Kuluev, 2018). SNPs are widely used for allelic polymorphism studies, seed purity testing, haplotype and pedigree analyses, as well as for genotyping and construction of genetic maps.

Obtaining SNP marker information is now possible for any plant at a whole genome scale through the use of next-generation high-throughput sequencing technologies. Identification of SNPs is possible using whole-genome sequencing (WGS) and genotyping by sequencing (GBS) strategies (Scheben et al., 2017). The aim of whole-genome sequencing is to obtain short random fragments (reads) of the whole genome DNA. This allows estimating DNA variation by aligning fragments to a reference genome or by genomic DNA *de novo* assembly. This can be challenging and expensive; price per genome exceeds $2000, also depending on the size and complexity of the genome, the desired level of completeness and computational resources (Narum et al., 2013). For example, sequencing a complete barley genome to the chromosome level costs around $60,000 (Monat et al., 2019). There are also specific methods of whole-genome sequencing with lower read depths that cost much less, $100–$400 per genome. However, according to the authors (Bimber et al., 2016), the accuracy of the resulting genotype data is decreased.

The genotyping by sequencing method is faster and more cost-efficient than the whole-genome sequencing method. For example, the cost of single barley genome sequencing by fragments in a GBS experiment does not exceed $30 (Monat et al., 2019). Two sequencing strategies can be applied in the GBS experiments. The first one uses site-specific restriction enzymes for fragmentation of DNA samples, after which sequencing of the resulting fragments is performed (Glaubitz et al., 2014). In the second method, unique adapter sequences are ligated to both ends of DNA fragments during library preparation (Elshire et al., 2011). Due to the fact that DNA fragments are only sequenced in the region of restriction sites, the GBS method does not sequence the full genome DNA sequence. This makes the sequencing process much cheaper. However, the number of SNPs that can be identified is lower than that obtained with whole-genome sequencing. Nevertheless, the amount of data obtained using the GBS method is sufficient to characterise the genetic diversity of agricultural plant populations with acceptable accuracy.

The GBS method has demonstrated its reliability and flexibility for a number of plant species and populations. GBS has been applied to the identification of molecular markers for genetic mapping (Poland et al., 2012), genomic selection (Poland et al., 2012), in genetic diversity studies (Lu et al., 2013; Peterson et al., 2014), variety identification (Wang et al., 2020; Rajendran et al., 2022), and studies in conservation biology and evolutionary ecology (Narum et al., 2013).

The GBS method significantly reduces the cost as well as the time required for sequencing the samples under study. This has led to the development of high-quality bioinformatics methods for the ever-expanding amount of sequenced data. To date, a number of bioinformatics pipelines for analyzing the data generated by GBS experiments have been developed. The workflows for existing pipelines of the GBS data analysis are similar and include raw reads preprocessing, data demultiplexing (if needed), mapping reads to a reference genome, SNP identification and analysis of genetic diversity.

The reads mapping step depends on the presence or absence of the reference genome sequence. In the first case, preprocessed reads are aligned to a reference genome using existing tools such as bowtie2 or bwa (Glaubitz et al., 2014; Torkamaneh et al., 2017; Wickland et al., 2017). In the absence of reference genome sequences, an additional step of "Mock Reference" sequence generation is applied (Melo et al., 2016). This method clusters reads by their similarity to identify consensus sequences (centroids) on the basis of which the fragments of the genome are assembled (Melo et al., 2016). These fragments of the genomic sequence are

used as the reference in subsequent analysis. Due to the similarity of the data processing steps, existing pipelines mainly differ in the software tools combined to perform the analysis. The combination should take into account various genomic characteristics, such as the number of polymorphisms detected, genome complexity, degree of heterozygosity, and the proportion of repetitive sequences in the whole genome. More advanced pipelines allow the selection of parameters for the organisms under study (Torkamaneh et al., 2017; Wickland et al., 2017), whereas earlier pipelines have some limitations. For example, TASSEL needs specification of the sequence length upper limit, which may result in the loss of a significant number of short raw reads (Glaubitz et al., 2014; Melo et al., 2016). Due to the ever-increasing number of sequenced libraries, pipelines must provide the capability to process a large amount of data in a single run. An important aspect of pipelines is the automation of the processing and the simplicity of the software installation.

In the present work, we have developed a GBS-DP bioinformatics pipeline for analysing GBS data. This pipeline incorporates the GBS data processing scheme proposed in (Jayakodi et al., 2020). The pipeline is applicable to any organism species. The pipeline can process large amounts of data (more than 400 samples) and is implemented using the Snakemake workflow system (Köster, Rahmann, 2012).

Materials and methods

Bioinformatics pipeline for analysing GBS data. The GBS-DP bioinformatics pipeline diagram is shown in Figure 1.

Pipeline input requires the path to the folder with the set of files with raw read sequences and the path to the reference genome file. Files with the read sequences should be in FASTQ format, reference genome, in FASTA format. If the libraries have barcode sequences, they must be demultiplexed beforehand by an external tool (the demultiplexing step is not included in GBS-DP).

The pipeline consists of three main steps: (1) data preprocessing, (2) polymorphism identification, (3) genetic diversity analysis. Data preprocessing includes quality control of raw reads, adapters removal, and construction of reference genome index files. Polymorphism identification includes mapping preprocessed reads to a reference genome, sorting the mapped reads, and searching for single nucleotide polymorphisms. Genetic diversity analysis is performed differently depending on whether the total size of files with polymorphism data exceeds 1 TB. Each stage is described in more detail below.

Data preprocessing. Quality control and adapter removal are performed by cutadapt (Martin, 2011). For the reads of each library, user should provide the list of adapter sequences in the pipeline configuration file.

The reference genome indexing is performed using the bwa tool ('index' option) (Li H., 2013).

Polymorphism identification. Mapping of preprocessed reads is performed by the bwa tool ('mem' option) (Li H., 2013) with the default parameters "–k 19 –w 100".

The mapping results are obtained in SAM format, converted into BAM format and sorted using samtools (Danecek et al., 2021) running 'view' and 'sort' options, respectively. Polymorphisms (SNPs, insertions and deletions (indels)) are identified using the sorted BAM by a combination of samtools ('mpileup' option) and bcftools ('call' option) (Danecek et al., 2021). It was previously shown using the wheat genome as an example (Yuao et al., 2020) that the

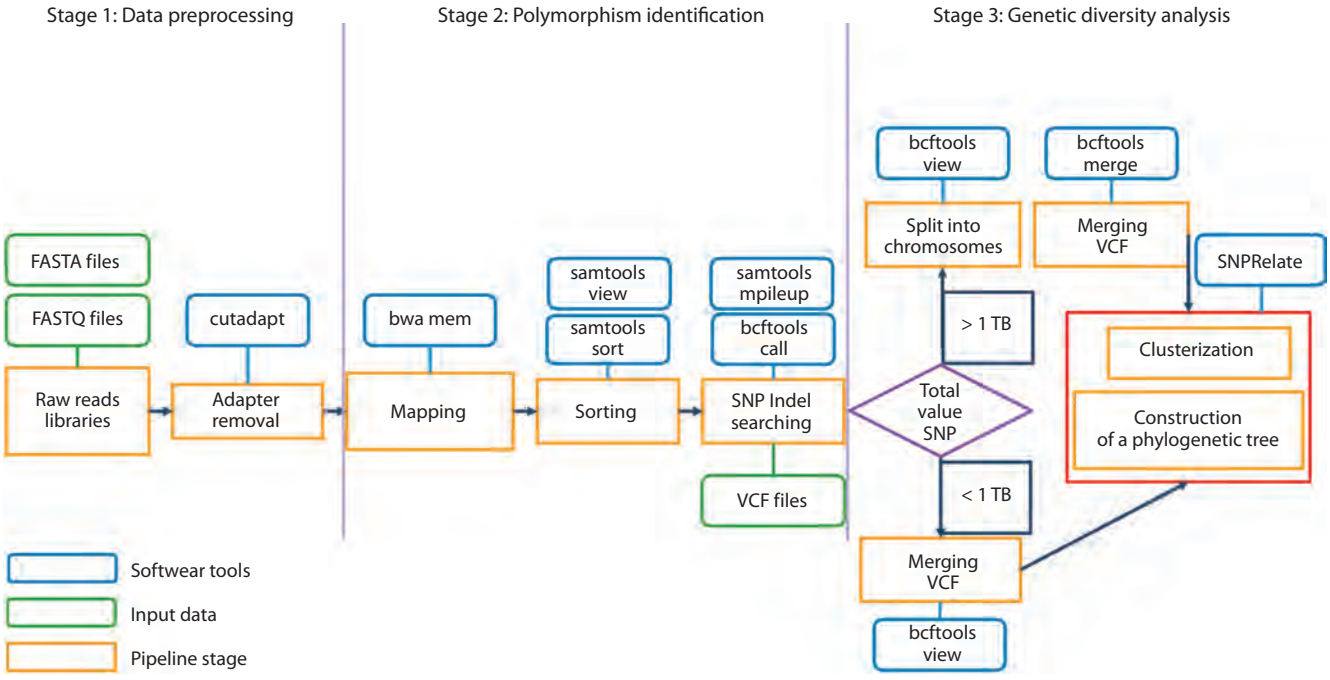


Fig. 1. The diagram of the GBS-DP bioinformatics pipeline.

samtools/mpileup + bwa-mem software combination used in our pipeline outperforms other combinations of polymorphism mapping and identification software.

Analysis of genetic diversity. The pipeline selects the way of genetic diversity analysis automatically depending on the total size of the VCF files obtained at the previous step.

The corresponding option is selected automatically and associated with increased load on the computer RAM when working with large data (if the total size of the received VCF files exceeds 1 TB). The processing option for data with the total volume less than 1 TB includes three steps. If the total size of files is lower than 1 TB, the pipeline performs the following steps:

1. VCF files containing information about polymorphisms for each sample are indexed using bcftools ('index' option) (Danecek et al., 2021).
2. The indexed files are merged into a single VCF file using bcftools ('merge' option). This file contains data on polymorphisms of all samples for all chromosomes.
3. The resulting file in VCF format is converted into GDS (Genomic Data Structure) format using the SeqArray package implemented in R (Zheng et al., 2017). This format allows significantly reducing the amount of RAM required for processing the results of polymorphism identification.

If the total size of VCF files is greater than 1 TB, the pipeline performs the following steps:

1. Each VCF file with polymorphism data for a specific sample is split by chromosome using bcftools ('view' option).
2. The resulting VCF files for each chromosome are indexed using bcftools ('index' option).
3. VCF files for each chromosome are merged for all samples. The resulting set of files represent the polymorphism data by chromosome for each sample.
4. VCF files for individual chromosomes are converted to GDS format. The resulting GDS files for each chromosome are then combined into a common GDS file using the snpgdsCombineGeno function of the SNPRelate package (Zheng et al., 2017).

The resulting polymorphism data merged from all samples are used for genetic diversity. It should be noted that important information about SNP distribution in the genomic sequence is represented by the linkage disequilibrium (LD) parameter (Ponomarenko, 2018). Two alleles of different loci are in linkage disequilibrium when the frequency of the haplotype comprising them differs significantly from the frequency expected under random segregation (Gabriel et al., 2002). The value of LD depends on a number of factors: the magnitude and rate of gene drift, genetic admixtures in the population, mutations and recombinations, and population size (Aulchenko, Aksenovich, 2006). LD is usually estimated by the linkage disequilibrium coefficient (D), but this measure is not always convenient because the range of its possible values depends on the frequencies of the alleles to which it refers. This makes it difficult to compare the level of linkage disequilibrium between different pairs of alleles. Thus, the D coefficient is normalised on the basis of the Pearson correlation coefficient r^2, which varies from 0 to 1. The closer the value of r^2 is to 0, the more likely it is that the identified SNPs are random.

The LD parameter is estimated by the GBS-DP pipeline using the merged file containing polymorphism information for all libraries across all chromosomes in GDS format. The R package SNPRelate (Zheng et al., 2017), function snpgdsLDpruning, is used for LD estimation.

Additionally, principal component analysis is applied for filtered SNPs, which is performed using the R package SNPRelate. The SNPRelate package is also used to build a phylogenetic tree using hierarchical clustering method.

System requirements and installation. The GBS-DP pipeline is implemented using the Snakemake v6.0.0 workflow management system (Köster, Rahmann, 2012), a tool for creating data analysis pipelines implemented in Python. Pipelines created in this environment can be easily scaled for server, cluster, network and cloud environments. Snakemake is compatible with the Conda system, making it easy to install new programs required for the pipeline. The pipeline is designed for the Linux operating system. It requires a minimum of 10 GB of RAM to run (the more data, the more RAM needed). To run the pipeline, user needs to specify parameters in the configuration file. The code and step-by-step instructions for running the pipeline are available at https://github.com/artempronozin95/GBS-DP-bioinformatics-pipeline-for-genotyping-by-sequencing-data-processing/tree/main.

Data for test analysis. For the test application of the GBS-DP pipeline in the present work, we used project PRJEB39633 from the European Nucleotide Archive (ENA) database (Leinonen et al., 2010), which contains GBS sequencing data for a barley (*Hordeum vulgare*) population derived from a cross between the six-row barley variety Morex and the mutant line *luteostrians*-P1 (*lst/LST*) (Li M. et al., 2021). Libraries were obtained using a combination of MspI and PstI restriction enzymes (Wendler et al., 2015). In total, the PRJEB39633 project contains 679 libraries for 272 genotypes; there is an average of 3 libraries per genotype, so library reads for the same genotype were combined before analysis.

We used the *H. vulgare* reference genome v. 51 sequence (IBSC_v2) downloaded from the Ensembl plants database (Bolser et al., 2016).

Results

Supplementary Material[1] demonstrates the processing time at different stages of the GBS-DP pipeline execution for different numbers of barley libraries (10, 50, 100, 150, 200 and 272). The characteristics of the computational node are as follows: AMD EPYC 74521 processor, 32 cores, 1 TB memory capacity. For the analysis, we used 100 GB of RAM and 20 processor cores. The longest time was spent on generating a merged file containing polymorphisms. However, it can be seen that the time taken to generate a merged file for 200 libraries is lower than that for 150 libraries; this is due to the usage of big data processing mode, which speeds up the calculation process.

The pipeline provides the results of the basic evaluation of sequenced libraries. The read length for each library is 107 nt. The average read depth (Fig. 2, *a*) ranges from 2–8,

[1] Supplementary Material is available at:
http://vavilov.elpub.ru/jour/manager/files/Suppl_Pronozin_Engl_27_7.pdf

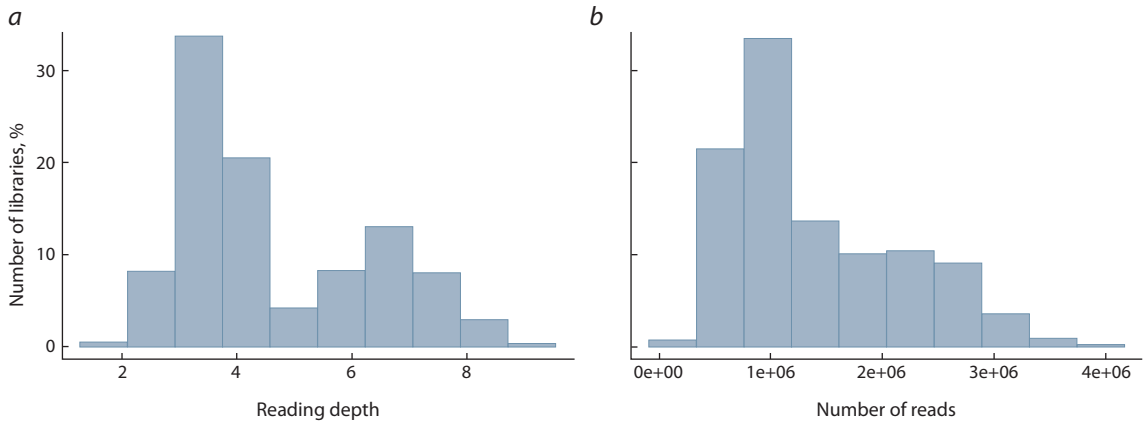


Fig. 2. Distribution of average depth of reads mapping (*a*) and distribution of mapped reads number for libraries (*b*).

which is an acceptable value for the GBS method. More than 30 % of the libraries contain more than 1,000,000 reads (see Fig. 2, *b*). On average for one library, the coverage of the barley reference genome (4,225,577,519 nt) with DNA fragments is 3 % of the total length.

The pipeline also provides the results of the search for polymorphisms between the investigated genotypes. For the 272 samples analysed, 447,409 SNPs were identified. The total number of indels is 46,557. The median value of transitions/transversions = 1.75, indicating the predominance of transitions. The estimate of the LD parameter (r^2) is 0.5. After applying the LD filter, 45,402 polymorphic and independent SNPs remained.

The distribution of the detected SNPs by chromosome showed that more SNPs were detected for chromosomes 3, 6 and 7 (Fig. 3). The main results of the pipeline are principal component analysis of genotypes based on the detected SNPs (Fig. 4) and construction of a phylogenetic tree. The results of principal component analysis based on 45,402 SNPs allow identifying three distinct clusters within the population. They are clearly distinguished in the scatter plot in the space of the first two components (see Fig. 4). However, the total proportion of variance attributable to these two components is small (20 %), which may indicate an overall high level of genetic diversity in the obtained plant population.

The phylogenetic tree constructed by the hierarchical clustering method is shown in Figure 5. Samples in the tree diagram are colored by cluster membership (see Fig. 4). It allows us to identify three large clusters in the population, which is consistent with the data presented in Figure 4.

Discussion

The decreasing cost and time required for GBS sequencing has led to a large number of experiments performed by this method. For example, the IPK Gatersleben barley genetic profile database (Milner et al., 2019) contains 22,626 samples obtained by the GBS method. Such a large number of samples requires a fast and high-quality data processing method. To date, pipelines processing GBS results already exist. However, despite the qualitatively selected software packages and the possibility to adjust parameters to the organisms under study, these pipelines have some disadvantages. For example, GBS-SNP-CROP and TASSEL have no possibility to automate the calculation process (each step should be started manually), which significantly reduces the speed of the study. GB-eaSy does not allow simultaneous research of several libraries of raw reads at once. In all existing pipelines, there is no possibility to switch off an unnecessary or passed step. For example, if there is no way to provide barcode data for the libraries being examined, then none of the listed pipelines will work. Also, in most pipelines, there is no possibility of automatic installation of all necessary software packages.

The pipeline we developed is based on the method proposed in (Jayakodi et al., 2020). In this paper, the bioinformatics tools are selected in such a way as to provide the most accurate polymorphism search result. However, this method is well applicable for small data, up to 50 libraries; as the number of libraries increases, the load on RAM and the space occupied on the hard disk increases. This leads to errors and interruption of the computation process. Thus, we proposed an approach for large GBS data processing based on (Jayakodi et al., 2020) method. The results of this approach are summarised in Supplementary Material and Figure 6.

As can be seen from Figure 6, our proposed approach significantly speeds up the calculation process for large data, but for small data, the difference in calculation speed is not large. Therefore, this mode is activated only for the VCF file data with the total size exceeding 500 GB.

The proposed pipeline was built using the Snakemake workflow manager. This method of implementation allows to automatically take into account the completed tasks for each sample, which eliminates the duplication of tasks, and also allows to resume the calculation process from the moment of its last interruption (for example, due to an error). Modular structure allows for more convenient functionality of manipulation of the pipeline steps (exclusion, addition, switching off some steps). Snakemake also has the ability to automatically install all the necessary software for the pipeline.

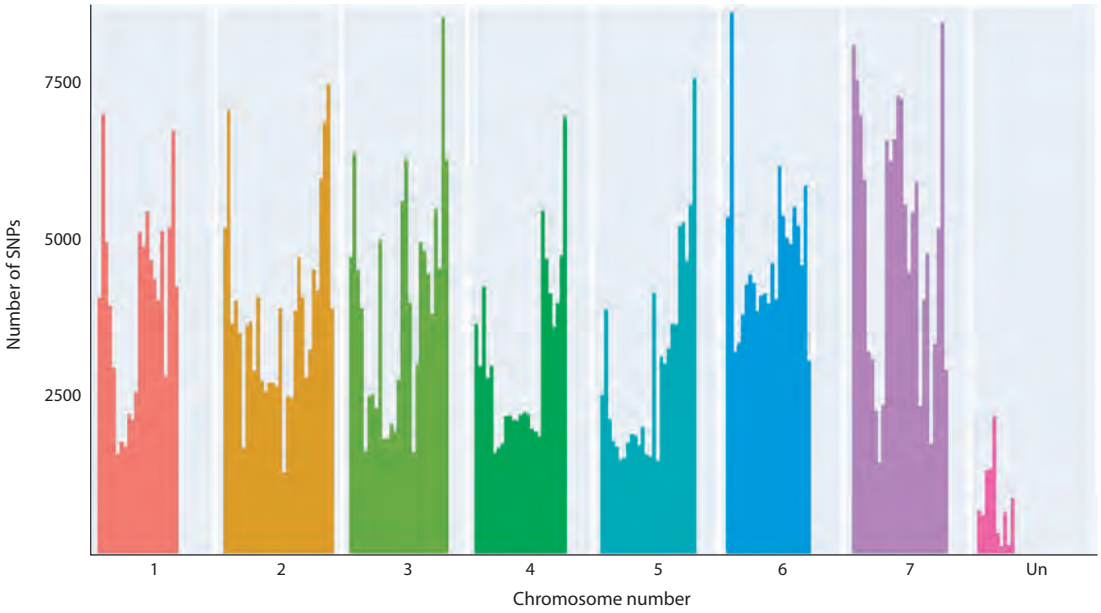


Fig. 3. Distribution of detected SNPs in the barley genome.

X axis is the coordinates of SNPs on chromosomes, *Y* axis is the number of SNPs corresponding to these coordinates.

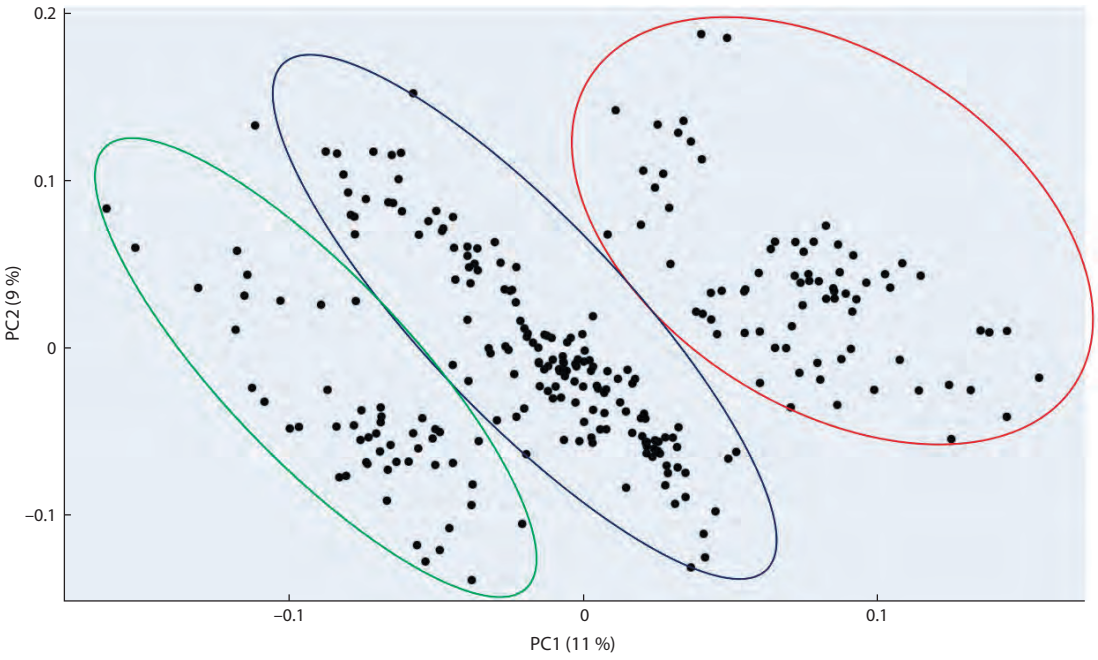


Fig. 4. Genotype scatter diagram for the barley population resulting from the cross between the variety Morex and the mutant line *luteostrians*-P1 (*lst/LST*) for the two principal components obtained from the genetic diversity analysis by the GBS-DP pipeline.

The proportion of the total variance is given in parentheses next to the component names.

Conclusion

Genotyping by sequencing methods have demonstrated their reliability and flexibility for a number of plant species and populations. They have reduced both the cost and the time required to sequence the samples under study, which has allowed even more sequencing to be performed. In this work, we proposed a GBS-DP bioinformatics pipeline, which allows us to process large-scale sequencing data performed by the GBS method. The results demonstrate a fairly high speed of this pipeline for both large data (more

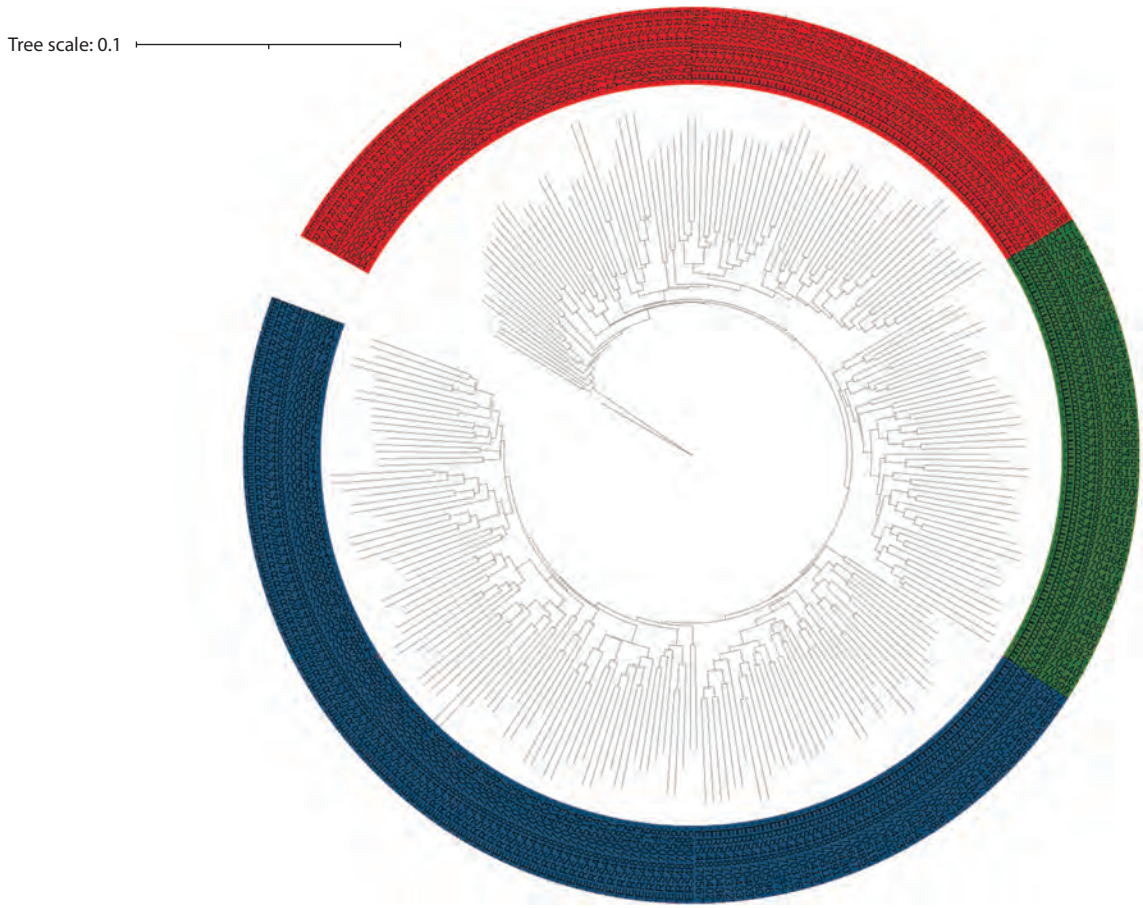


Fig. 5. Phylogenetic tree of 272 barley samples constructed by the hierarchical clustering method by their genetic similarity estimated using the GBS data.

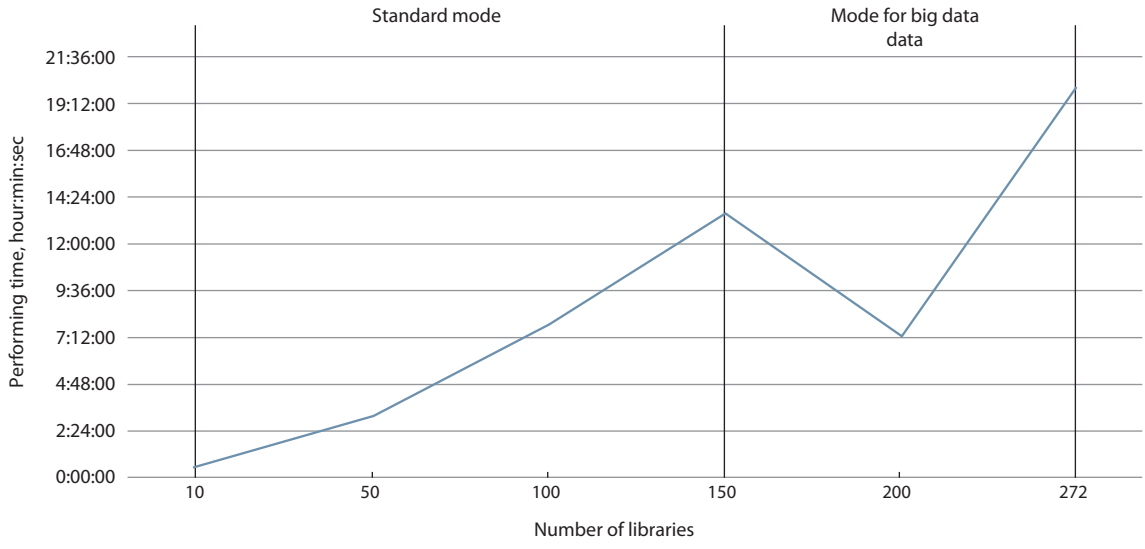


Fig. 6. Dependence of time spent on the conveyor operation on the number of libraries under study.

than 400 libraries) and small data (~30 libraries). The pipeline also provides analysis of detected polymorphisms.

References

Aulchenko Yu.S., Aksenovich T.I. Methodological approaches and strategies for mapping genes controlling complex human traits. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2006;10(1):189-202 (in Russian)

Bimber B.N., Raboin M.J., Letaw J., Nevonen K.A., Spindel J.E., McCouch S.R., Cervera-Juanes R., Spindel E., Carbone L., Ferguson B., Vinson A. Whole-genome characterization in pedigreed non-human primates using genotyping-by-sequencing (GBS) and imputation. *BMC Genomics*. 2016;17(1):676. DOI 10.1186/s12864-016-2966-x

Bolser D., Staines D.M., Pritchard E., Kersey P. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards D. (Ed.) Plant Bioinformatics. Methods in Molecular Biology. Vol. 1374. New York: Humana Press, 2016;115-140. DOI 10.1007/978-1-4939-3167-5_6

Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., Li H. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2): giab008. DOI 10.1093/gigascience/giab008

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6(5): e19379. DOI 10.1371/journal.pone.0019379

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., Altshuler D. The structure of haplotype blocks in the human genome. *Science*. 2002;296(5576):2225-2229. DOI 10.1126/science.1069424

Glaubitz J.C., Casstevens T.M., Lu F., Harriman J., Elshire R.J., Sun Q., Buckler E.S. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*. 2014;9(2):e90346. DOI 10.1371/journal.pone.0090346

Jayakodi M., Padmarasu S., Haberer G., Bonthala V.S., Gundlach H., Monat C., Lux T., Kamal N., Lang D., Himmelbach A., Ens J., Zhang X.Q., Angessa T.T., Zhou G., Tan C., Hill C., Wang P., Schreiber M., Boston L.B., Plott C., Jenkins J., Guo Y., Fiebig A., Budak H., Xu D., Zhang J., Wang C., Grimwood J., Schmutz J., Guo G., Zhang G., Mochida K., Hirayama T., Sato K., Chalmers K.J., Langridge P., Waugh R., Pozniak C.J., Scholz U., Mayer K.F.X., Spannagl M., Li C., Mascher M., Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. 2020;588(7837): 284-289. DOI 10.1038/s41586-020-2947-8

Kanukova K.R., Gazaev I.Kh., Sabanchieva L.K., Bogotova Z.I., Appaev S.P. DNA markers in crop production. *Izvestiya Kabardino-Balkarskogo Nauchnogo Tsentra RAN = News of the Kabardin-Balkar Scientific Center of RAS.* 2019;6(92):220-232. DOI 10.35330/1991-6639-2019-6-92-220-232 (in Russian)

Khlestkina E.K. Molecular markers in genetic studies and breeding. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):1044-1054 (in Russian)

Köster J., Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. DOI 10.1093/bioinformatics/bts480

Leinonen R., Akhtar R., Birney E., Bower L., Cerdeno-Tárraga A., Cheng Y., Cleland I., Faruque N., Goodgame N., Gibson R., Hoad G., Jang M., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Sobhany S., Ten Hoopen P., Vaughan R., Zalunin V., Cochrane G. The European nucleotide archive. *Nucleic Acids Res.* 2011; 39(Database issue):D28-D31. DOI 10.1093/nar/gkq967

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. 2013. DOI 10.48550/arXiv.1303.3997

Li M., Guo G., Pidon H., Melzer M., Prina A.R., Börner T., Stein N. ATP-dependent *Clp* protease subunit *C1, HvClpC1,* is a strong candidate gene for barley variegation mutant *luteostrians* as revealed by genetic mapping and genomic re-sequencing. *Front. Plant Sci.* 2021;12:664085. DOI 10.3389/fpls.2021.664085

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 2013;9(1):e1003215. DOI 10.1371/journal.pgen.1003215

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17(1):10-12. DOI 10.14806/ej.17.1.200

Melo A.T., Bartaula R., Hale I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*. 2016;17(1):29. DOI 10.1186/s12859-016-0879-y

Milner S.G., Jost M., Taketa S., Mazón E.R., Himmelbach A., Oppermann M., Weise S., Knüpffer H., Basterrechea M., König P., Schüler D., Sharma R., Pasam R.K., Rutten T., Guo G., Xu D., Zhang J., Herren G., Müller T., Krattinger S.G., Keller B., Jiang Y., González M.Y., Zhao Y., Habekuß A., Färber S., Ordon F., Lange M., Börner A., Graner A., Reif J.C., Scholz U., Mascher M., Stein N. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 2019;51(2):319-326. DOI 10.1038/s41588-018-0266-x

Monat C., Schreiber M., Stein N., Mascher M. Prospects of pan-genomics in barley. *Theor. Appl. Genet.* 2019;132(3):785-796. DOI 10.1007/s00122-018-3234-z

Narum S.R., Buerkle C.A., Davey J.W., Miller M.R., Hohenlohe P.A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* 2013;22(11):2841-2847. DOI 10.1111/mec.12350

Peterson G.W., Dong Y., Horbach C., Fu Y.-B. Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. *Diversity*. 2014;6(4):665-680. DOI 10.3390/d6040665

Poland J., Endelman J., Dawson J., Rutkoski J., Wu S., Manes Y., Dreisigacker S., Crossa J., Sánchez-Villeda H., Sorrells M., Jannink J.-L. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*. 2012;5(3):103-113. DOI 10.3835/plantgenome2012.06.0006

Ponomarenko I.V. Selection of polymorphic loci for association analysis in genetic-epidemiological studies. *Nauchnye Rezultaty Biomeditsynskikh Issledovaniy = Research Results in Biomedicine*. 2018;4(2):40-54. DOI 10.18413/2313-8955-2018-4-2-0-5 (in Russian)

Rajendran N.R., Qureshi N., Pourkheirandish M. Genotyping by sequencing advancements in barley. *Front. Plant Sci.* 2022;13:931423. DOI 10.3389/fpls.2022.931423

Scheben A., Batley J., Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.* 2017;15(2):149-161. DOI 10.1111/pbi.12645

Sukhareva A.S., Kuluev B.R. DNA markers for genetic analysis of crops. *Biomika = Biomics*. 2018;10(1):69-84. DOI 10.31301/2221-6197.bmcs.2018-15 (in Russian)

Torkamaneh D., Laroche J., Bastien M., Abed A., Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*. 2017;18(1):5. DOI 10.1186/s12859-016-1431-9

Wang N., Yuan Y., Wang H., Yu D., Liu Y., Zhang A., Gowda M., Nair S.K., Hao Z., Lu Y., San Vicente F., Prasanna B.M., Li X., Zhang X. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 2020;10(1):16308. DOI 10.1038/s41598-020-73321-8

Wendler N., Mascher M., Himmelbach A., Johnston P., Pickering R., Stein N. Bulbosum to go: a toolbox to utilize *Hordeum vulgare/bul-*

bosum introgressions for breeding and beyond. *Mol. Plant*. 2015; 8(10):1507-1519. DOI 10.1016/j.molp.2015.05.004

Wickland D.P., Battu G., Hudson K.A., Diers B.W., Hudson M.E. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*. 2017;18:586. DOI 10.1186/s12859-017-2000-6

Yao Z., You F.M., N'Diaye A., Knox R.E., McCartney C., Hiebert C.W., Pozniak C., Xu W. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics*. 2020;21(1):360. DOI 10.1186/s12859-020-03704-1

Zheng X., Gogarten S.M., Lawrence M., Stilp A., Conomos M.P., Weir B.S., Laurie C., Levine D. SeqArray – a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*. 2017;33(15):2251-2257. DOI 10.1093/bioinformatics/btx145

ORCID ID
A.Yu. Pronozin orcid.org/0000-0002-3011-6288
E.A. Salina orcid.org/0000-0001-8590-847X
D.A. Afonnikov orcid.org/0000-0001-9738-1409

Acknowledgements. The work was supported by the budget project FWNR-2022-0020.

Transparency of financial activities. The authors have no financial interest in the materials or methods presented.

Conflict of interest. The authors declare no conflict of interest.

Received July 21, 2023. Revised September 8, 2023. Accepted September 9, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):746-754
DOI 10.18699/VJGB-23-87

Original Russian text https://vavilovj-icg.ru/

The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis

T.A. Bukharina[1, 2] ✉, V.P. Golubyatnikov[3], D.P. Furman[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ bukharina@bionet.nsc.ru

Abstract. Identification of the mechanisms underlying the genetic control of spatial structure formation is among the relevant tasks of developmental biology. Both experimental and theoretical approaches and methods are used for this purpose, including gene network methodology, as well as mathematical and computer modeling. Reconstruction and analysis of the gene networks that provide the formation of traits allow us to integrate the existing experimental data and to identify the key links and intra-network connections that ensure the function of networks. Mathematical and computer modeling is used to obtain the dynamic characteristics of the studied systems and to predict their state and behavior. An example of the spatial morphological structure is the Drosophila bristle pattern with a strictly defined arrangement of its components – mechanoreceptors (external sensory organs) – on the head and body. The mechanoreceptor develops from a single sensory organ parental cell (SOPC), which is isolated from the ectoderm cells of the imaginal disk. It is distinguished from its surroundings by the highest content of proneural proteins (ASC), the products of the *achaete-scute* proneural gene complex (*AS-C*). The SOPC status is determined by the gene network we previously reconstructed and the *AS-C* is the key component of this network. *AS-C* activity is controlled by its subnetwork – the central regulatory circuit (CRC) comprising seven genes: *AS-C*, *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), and *extramacrochaete* (*emc*), as well as their respective proteins. In addition, the CRC includes the accessory proteins Daughterless (DA), Groucho (GRO), Ubiquitin (UB), and Seven-in-absentia (SINA). The paper describes the results of computer modeling of different CRC operation modes. As is shown, a cell is determined as an SOPC when the ASC content increases approximately 2.5-fold relative to the level in the surrounding cells. The hierarchy of the effects of mutations in the CRC genes on the dynamics of ASC protein accumulation is clarified. *AS-C* as the main CRC component is the most significant. The mutations that decrease the ASC content by more than 40 % lead to the prohibition of SOPC segregation.
Key words: central regulatory circuit; gene network; mathematical model; computer modeling; drosophila; *achaete-scute* complex; mutations.

For citation: Bukharina T.A., Golubyatnikov V.P., Furman D.P. The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):746-754. DOI 10.18699/VJGB-23-87

Центральный регуляторный контур генной сети морфогенеза механорецепторов дрозофилы: анализ *in silico*

Т.А. Бухарина[1, 2] ✉, В.П. Голубятников[3], Д.П. Фурман[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ bukharina@bionet.nsc.ru

Аннотация. Выявление механизмов генетического контроля формирования пространственных структур остается одной из актуальных задач биологии развития. Для ее решения используются как экспериментальные, так и теоретические подходы и методы, в том числе методология генных сетей, а также методы математического и компьютерного моделирования. Реконструкция и анализ генных сетей, обеспечивающих становление признака, позволяют интегрировать существующие экспериментальные данные, выявить ключевые звенья и внутрисетевые связи, обеспечивающие функционирование сетей. Для получения динамических характеристик исследуемых систем, предсказания их состояния и поведения привлекаются методы математического и компьютерного моделирования. Одним из примеров пространственной морфологической структуры является щетиночный рисунок дрозофилы со строго определенным расположением на голове и теле мухи его составляющих – механорецепторов (внешних сенсорных органов). Механорецептор развивается из единственной родительской клетки (PKCO),

которая выделяется из клеток эктодермы имагинального диска. Ее отличает от окружения наибольшее содержание пронейральных белков (ASC) – продуктов комплекса пронейральных генов *achaete-scute* (*AS-C*). Статус РКСО обеспечивается реконструированной нами ранее генной сетью, ключевым объектом которой является комплекс генов *AS-C*. Контроль активности комплекса осуществляется ее подсетью – центральным регуляторным контуром в составе семи генов (*AS-C*, *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), *extramacrochaete* (*emc*)) и одноименных белков. Кроме того, в состав центрального регуляторного контура входят вспомогательные белки Daughterless (DA), Groucho (GRO), Ubiquitin (UB) и Seven-in-absentia (SINA). В работе приведены результаты компьютерного моделирования различных режимов функционирования контура. Показано, что клетка детерминируется как РКСО при повышении содержания ASC примерно в два с половиной раза относительно уровня в клетках окружения. Выявлена иерархия влияния мутаций в генах контура на динамику накопления белков ASC. Наиболее значим главный компонент центрального регуляторного контура – *AS-C*. Мутации, снижающие содержание ASC более чем на 40 %, приводят к запрету выделения родительской клетки сенсорного органа.
Ключевые слова: центральный регуляторный контур; генная сеть; математическая модель; компьютерное моделирование; дрозофила; *achaete-scute* комплекс; мутации.

Introduction

The current views on the control of biological processes, including cell differentiation, growth and development of organisms, and construction of spatial structures, are united in the concept of gene networks. According to this concept, gene networks (GNs) are the molecular genetic systems that provide the formation of all phenotypic characteristics of organisms (molecular, biochemical, structural, morphological, ethological, physiological, cognitive, and so on) based on the information coded for in their genomes. Kolchanov et al. (2013) define GNs as the groups of concertedly operating genes that interact with one another via both their primary products (RNAs and proteins) and the diverse metabolites and other secondary products of GN operation.

The GNs are reconstructed based on the analysis of experimental data, which gives both the most comprehensive and systematized description of a considered biological system or a process (Schlitt et al., 2003; Zhu et al., 2007; Emmert-Streib, Glazko, 2011; Chasman et al., 2016). An important feature of the GNs is regulatory circuits, which ensure their correct function and implementation of the program that forms a phenotypic trait.

Mathematical and computer modeling makes it possible to acquire the most comprehensive insight into the GN arrangement and behavior and is widely used to clarify the structure–function organization of GNs, architecture of their inner links, detection of the key elements and modules, and patterns of their operation and evolution.

The GNs "Neurogenesis:prepattern", "Neurogenesis:determination", and "Neurogenesis:asymmetric division", which we have earlier reconstructed are examples of the networks responsible for the development of ordered structures during ontogenesis. Together, these GNs provide a definite composition of mechanoreceptors (sensory organs of the peripheral nervous system) on the head and body of drosophila (Furman, Bukharina, 2022). Analysis of these networks has elicited the most important connecting link that controls their operation, namely, central regulatory circuit (CRC). It is a correct CRC operation in the "Neurogenesis:determination" GN that determines the implementation of the key event in the morphogenesis of each mechanoreceptor – the definition of a single sensory organ parental cell (SOPC), which is separated within a proneural cluster, the group of epidermal cells within imaginal disk (Furman, Bukharina, 2022). The parental cell differs from the surrounding ones by the content of proneural ASC proteins, coded for by the gene complex of the same name, *achaete-scute complex* or *AS-C* (Reeves, Posakony, 2005). An increased ASC content is the factor that determines the neural fate of a cell. By ensuring the development of both individual mechanoreceptors and their overall array, the so-called bristle pattern, the CRC regulates the production of these proteins to the level necessary for a cell to acquire an SOPC status (Furman, Bukharina, 2022).

Although the morphogenesis of mechanoreceptors has been long studied, it is unfortunately still far from an exhaustive description. It is only qualitatively characterized: the players in this process (genes and proteins) are known and the general concept of their interaction is formed; however, most of the quantitative parameters as well as a relative contribution of the involved genes have not been experimentally determined. Note that the scientists studying biological systems often encounter the situation of data incompleteness; here, mathematical and computer modeling is the tool allowing this problem to be resolved. A model with adequately selected parameters makes it possible not only to assess the current state of a system or an ongoing process, but also has a predictive value. Numerical experiments conducted with the help of mathematical models allow potential operation modes of a system to be examined, its future states to be forecasted, and its new functions to be predicted by changing parameters or adding new assumptions. In many cases, modeling is the only way to understand the processes taking place in a system when their characteristics cannot be directly measured in a biological experiment.

Modeling of the morphogenesis of mechanoreceptors at the stage of SOPC segregation from the cells of proneural cluster has been earlier attempted; however, the authors confined themselves to integrated characteristics and general schemes of intracellular and intercellular interactions of gene groups without (or with minimum) detailing of their composition and particular contributions of individual players (Marnellos, Mjolsness, 1998; Meir et al., 2002; Ghysen, Thomas, 2003; Hsu et al., 2006; Corson et al., 2017; Yasugu, Sato, 2022). Any integral concept of the mechanisms underlying the intracellular interactions in SOPC formation is still absent, as well as the quantitative characteristics for the content of ASC proteins critical for determining the neural fate of a cell are not determined and the degree of the influence of CRC components on the expression of *AS-C* genes is vague.

The goal of this work was to construct a mathematical model of CRC operation taking into account the roles of the

constituent genes that would comprehensively describe the intracellular events in presumptive SOPC determining the dynamics of ASC content and to perform the computer experiments for verifying the model stability and its compliance with experimental data.

Materials and methods

Object of modeling is the CRC (see Fig. 1 for the scheme). In addition to the *AS-C* proneural genes and the ASC proteins they code for, the circuit comprises the genes *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), and *extramacrochaete* (*emc*) and the corresponding proteins. The CRC also contains the proteins Daughterless (DA), Groucho (GRO), Ubiquitin (UB), and Seven-in-absentia (SINA). All components are connected with *AS-C* via activation–repression interactions.

The content of proneural ASC proteins in SOPC is determined via auto- and trans-regulation of *AS-C* gene activity. The activating autoregulation is implemented by the ASC/DS heterodimers and the repression, by ASC/EMC heterodimers. The trans-regulation of the CRC genes with an activating effect is performed by the Senseless and Charlatan proteins and with a negative effect, by the Hairy/GRO and ASC/EMC complexes (Cabrera, Alonso, 1991; Van Doren et al., 1992, 1994; Cabrera et al., 1994; Vaessin et al., 1994; Nolo et al., 2000; Escudero et al., 2005) (see Fig. 1).

Certain additional mechanisms make it possible to avoid the repressive effect of Hairy/GRO and ASC/EMC on *AS-C*. In particular, the activation of gene *scratch* by the ASC/DA heterodimers entails the repression of *hairy* transcriptional activity (Roark et al., 1995) and, as a consequence, an increase in the expression of *AS-C*. The activation of the *chn* gene represses the transcription of *hairy* and *emc* (Yamasaki et al., 2011) and leads to the same effect, that is, an increase in the *AS-C* expression (see Fig. 1).

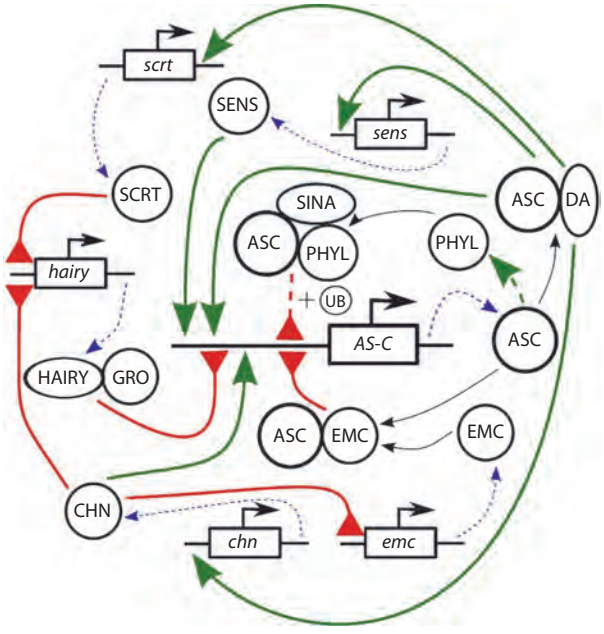


Fig. 1. Scheme of the central regulatory circuit of the gene networks underlying the development of drosophila macrochaetes: *AS-C*, *achaete-scute* gene complex; ASC, achaete-scute complex proteins; da, daughterless; gro, groucho; sens, senseless; emc, extramacrochaete; chn, charlatan; and scrt, scratch.

Green arrows show activator effects (solid line, direct and dashed, mediated) and red arrows with chopped ends denote repressor effects (solid line, direct and dashed, mediated). The earlier published scheme (Golubyatnikov et al., 2015) has been updated by adding the ASC protein degradation system.

Expression of the *sens*, *scrt*, and *chn* genes and, thus, the production of the corresponding proteins are regulated by the ASC/DA heterodimers, which initiate their transcription (Cabrera, Alonso, 1991; Vaessin et al., 1994; Nolo et al., 2000; Escudero et al., 2005) (see Fig. 1).

The CRC operation also requires the players involved in protein degradation, namely, ubiquitin (UB) and the E3 ubiquitin ligase Seven-in-absentia (SINA), as well as the adaptor protein Phyllopod (PHYL) (Pi et al., 2001; Chang et al., 2008).

Model. The proposed dynamical model of *AS-C* activity is described with a system of ordinary differential equations (1) (Bukharina et al., 2020):

$$\begin{aligned}
\frac{dx}{dt} &= k_x \frac{\sigma_1(D\cdot x)+\sigma_4(z)+\sigma_6(w)}{(1+G\cdot y)(1+E\cdot x)} - (1+p(t-\tau)\cdot U\cdot S)m_x\cdot x,\\
\frac{dy}{dt} &= k_y \frac{C_y}{(d_1+u)(d_2+w)} - m_y\cdot y,\\
\frac{dE}{dt} &= k_e \frac{C_e}{(d_3+w)(d_2+w)} - m_e\cdot E,\\
\frac{dz}{dt} &= k_z s_4(D\cdot x) - m_z\cdot z,\\
\frac{du}{dt} &= k_u s_5(D\cdot x) - m_u\cdot u,\\
\frac{dw}{dt} &= k_w s_6(D\cdot x) - m_w\cdot w,\\
\frac{dp}{dt} &= k_p \frac{s_7(D\cdot x)\cdot h(t-\tau)\cdot(t-\tau)^2}{(L+h(t-\tau)\cdot(t-\tau)^2)(1+G\cdot y)(1+E\cdot x)} - m_p\cdot p.
\end{aligned} \qquad (1)$$

The variables in this system are the concentrations of the CRC proteins in the cell: $x(t)$ is the content of ASC; $y(t)$, of Hairy; $E(t)$, of Extramacrochaete; $z(t)$, $u(t)$, $w(t)$, and $p(t)$, the concentrations of Senseless, Scratch, Charlatan, and Phyllopod, respectively.

To take into account the mutations of the genes that compose the CRC, the model contains non-negative coefficients k_x, k_y, k_e, and so on reflecting the degrees of influence of the mutations on the synthesis of the corresponding proteins. The values of these coefficients do not exceed unity; $k = 1$ corresponds to the normal operation of a gene; and $k = 0$ denotes a complete inactivation of a gene and the absence of the corresponding protein.

Parameters x_0, y_0, z_0, u_0, w_0, p_0, and E_0 denote the concentrations of the proteins ASC, Hairy, SENS, SCRT, CHN, PHYL, and EMC in the initial state of the CRC when the proneural cluster is already established, expression of all *AS-C* genes starts in all its cells, and all these cells still have equal neural potencies.

The values of parameters D, G, S, and U in system (1) are assumed constant since the concentrations of the corresponding proteins DA, GRO, SINA, and UB almost do not vary during the formation of parental cell. Parameters C_y, C_e, d_1, d_2, and d_3 are assumed constant as well.

Positive coefficients m_x, m_y, m_e, m_z, m_u, m_w, and m_p describe the degradation rates of the corresponding proteins.

The positive summand in the second equation of system (1) describes the negative feedbacks SCRT–Hairy and CHN–Hairy (see Fig. 1). The sigmoid functions σ_l, where $l = 1, 4, 6$ in the first equation of system (1), and the sigmoid functions s_i, where $i = 4, 5, 6, 7$ in the fourth–seventh equations of system (1), correspond to the positive feedbacks shown in Figure 1 with green arrows:

$$\sigma_l(q) = \frac{a_l q^{n_l}}{b_l + q^{n_l}},$$

$$s_i(q) = \frac{\alpha_i q^{\nu_i}}{\beta_i + q^{\nu_i}}.$$

Here, α_i, β_i, ν_i and a_l, b_l, n_l are positive parameters, $q \geq 0$ (Bukharina et al., 2015).

The model anticipates the choice of the CRC operation lifetime (T) and the moment (τ) when protein PHYL appears in the cell. The CRC functions until the cell starts to divide; hence, time T directly depends on τ: the later PHYL appears, the later the cell divides and the longer the CRC continues its operation. In the equation with delay, the function $p(t_0)$ is taken equal to 0 for $0 \leq \tau \leq t$.

Software. A special program complex based on the Shiny package has been designed for the numerical experiments with the CRC model described above and visualization of their results. The software makes it possible to elaborate interactive web applications with graphical user interface with the help of the R language (https://shiny.rstudio.com/).

The developed web application (https://gene-nets-simulation.shinyapps.io/crc-asc-modeler/) allows the CRC operation modes to be simulated for different values of the parameters of system (1) and the results of these numerical experiments to be visualized as plots. Here, the parameters of the system are chosen in accordance with the results of biological experiments.

Results and discussion

Let us consider the modeling results for different CRC operation modes.

Modeling of CRC operation in the presumptive parental cell of mechanoreceptor in the absence of any mutations in the constituent genes

Figure 2 shows the results of computer simulation of CRC operation in the future SOPC in the norm (absence of any mutations in the CRC constituent genes). The parameters of system (1) were selected taking into account the available published experimental data (Reeves, Posakony, 2005; Chang et al., 2008; Giri et al., 2022):

$$\begin{aligned}
&D = 1.6;\ G = 1;\ m_x = 0.3;\ U = 1.1;\ S = 5.5;\\
&a_1 = 2.9;\ n_1 = 1;\ b_1 = 1;\ a_4 = 5.8;\ n_4 = 1;\ b_4 = 5.6;\\
&a_6 = 6;\ n_6 = 1;\ b_6 = 5.7;\\
&C_y = 14.1;\ d_1 = 4.1;\ d_2 = 4.7;\ m_y = 0.5;\\
&C_e = 2.9;\ d_3 = 7.5;\ m_e = 0.4;\\
&\alpha_4 = 3;\ \nu_4 = 1.9;\ \beta_4 = 1.2;\ m_z = 1.6;\\
&\alpha_5 = 14.8;\ \nu_5 = 1.1;\ \beta_5 = 14.8;\ m_u = 2.3;\\
&\alpha_6 = 2;\ \nu_6 = 1;\ \beta_6 = 1;\ m_w = 1;\\
&\alpha_7 = 4.5;\ \nu_7 = 3.1;\ \beta_7 = 0.5;\ m_p = 0.6;\ L = 1.1;\\
&x_0 = 0.8;\ y_0 = 1.6;\ E_0 = 1.1;\ z_0 = 0.4;\ u_0 = 0;\ w_0 = 0;\ p_0 = 0;\\
&T = 28;\ \tau = 12,
\end{aligned} \qquad (2)$$

and coefficients $k = 1$ in all equations of system (1).

It is known that the SOPC determination for mechanoreceptors of different localizations takes different time (Cubas et al., 1991; Huang et al., 1991; Usui, Kimura, 1993). The time interval $T = 28$ h was selected as an interval close to the maximum necessary for determination of a neural cell fate (Huang et al., 1991). It is assumed that the CRC operation commences as early as the formation of proneural clusters 35–40 h before the puparium is formed when the expression of *AS-C* genes is first recorded (Cubas et al., 1991; Skeath, Carroll, 1991). The moment when proneural cluster is already formed, all its constituent cells display *AS-C* expression, and all of them still have equal neural potencies is regarded as the point zero.

The pattern of the changes in the content of ASC proteins in Figure 2 qualitatively matches the pattern observable in experiments (Reeves, Posakony, 2005; Chang et al., 2008). It is known that the content of ASC proteins gradually increases to reach a certain critical level after which the cell fate is unambiguously determined, namely, it becomes an SOPC. In the above-described numerical experiment, we got a smooth increase in the protein content over approximately 10 h to the level exceeding the initial one approximately 3.7-fold, that is, from 0.8 to 2.95.

Once the maximum is reached, the content of ASC proteins commences decreasing after a certain time interval to drop to almost zero value by the moment the SOPC starts dividing. This is determined by the switch-on of an additional regulatory mechanism associated with the degradation of ASC proteins (Chang et al., 2008). With the selected parameters, the model predicts that the ASC content commences to sharply decrease approximately in 15 h to reach the zero values during in 3 h.

It is important that the model excludes the possibility of any cyclic processes during the time interval limited by the moment of cell division, thereby demonstrating that the determination of a neural fate of the cell is irreversible. This also complies with the available published data (Reeves, Posakony, 2005; Chang et al., 2008).

According to different researchers, the SOPC segregation from proneural clusters for the mechanoreceptors of different localizations takes in the norm 9–12 to 28–30 h (Huang et al., 1991; Audibert et al., 2005; Kawamori et al., 2013). Note that the SOPC divisions for all mechanoreceptors are more or less

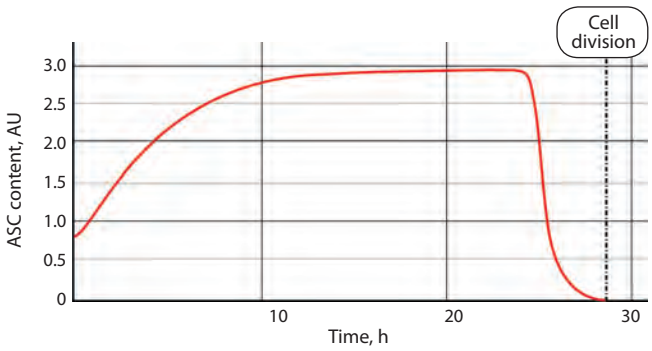


Fig. 2. Dynamics of ASC protein content in the mechanoreceptor presumptive parent cell in the norm (AU, arbitrary units).

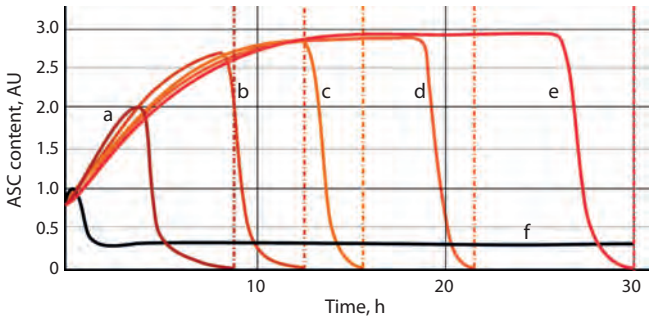


Fig. 3. Dynamics of the content of ASC proteins in the presumptive parent cell of mechanoreceptor for different time parameters.

(a–e) Parameter values are given in the text and (f) τ = 0 and *T* = 30 h. Vertical dashed lines denote cell division.

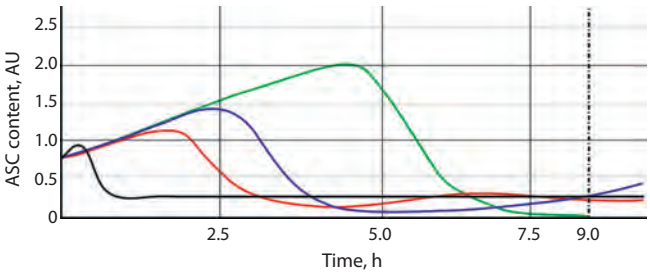


Fig. 4. Evaluation of the minimum level of ASC protein content in the presumptive SOPC sufficient for the cell to acquire a neural status.

See text for the values of time parameters.

synchronous and take place 0–3 h after pupation (Huang et al., 1991; Ayeni et al., 2016).

The first set of additional numerical experiments aimed at the testing of model stability to the change in time intervals required for the accumulation of ASC proteins in the amount necessary for the cell to achieve an SOPC status and to pass over to division (9 to 30 h). In this process, the value of parameter τ (the time moment when PHYL protein appears, which is critical for the transition of cell to division) was changed so that the values of parameter *T* (transition of SOPC to division) fall into the range of 9–30 h:

a) $T = 9$; $\tau = 2.1$;
b) $T = 18$; $\tau = 4$;
c) $T = 18$; $\tau = 6$;
d) $T = 22$; $\tau = 9$, and
e) $T = 28$; $\tau = 12$.

In additional experiments, the value of τ was taken to be 0 (that is, PHYL protein appeared simultaneously with ASC proteins) and parameter *T* was selected in an arbitrary manner to be 30 h or larger. The remaining parameters in these experiments remained constant and matched parameter set (2).

Figure 3 shows the plots illustrating the dynamics of protein contents in mechanoreceptor parental cell at the selected time parameters. As is evident, the patterns of plots (a–e), shown by different tints of red, are similar to one another and the plot in Figire 2. The curves differ only in the duration of the phase when the ASC content is at its maximum level. Note that the shape of the curve is retained in the selected range of τ and the corresponding *T* values, thereby demonstrating that the proposed model of CRC operation is stable. For the case of $\tau = 0$, which simulates the situation when PHYL (involved in the degradation of ASC proteins) appears without any delay, the shape of curve (f) in Figure 3, colored black, considerably differs from the remaining plots. The initial insignificant increase in the ASC content (not exceeding 16–17 % of the initial level) is followed by a decrease (to approximately half of the initial level) with subsequent plateau at a low level (although nonzero but insufficient for determining a cell as SOPC).

This result indirectly confirms the earlier assumption that a delayed appearance of the PHYL protein is the particular necessary condition for parent cell determination (Furman, Bukharina, 2022).

This model makes it possible to gain the insight into the dynamics of ASC content in presumptive SOPC. By varying parameter τ, it is possible to assess what is the minimum necessary and sufficient excess amount of ASC proteins in a cell as compared with the content in the surrounding cells that ensure a neural status. Here, it is necessary to take into account the experimentally determined fact that this process requires at least 9 h (Huang et al., 1991; Audibert et al., 2005; Kawamori et al., 2013). Figure 4 shows the modeling results for the τ values of 0 h (black plot), 0.5 h (red plot), 1 h (blue plot), and 2.1 h (green plot).

The value of $\tau = 2.1$ h is the first one when two conditions for cell transition to division are fulfilled: (1) the content of ASC proteins has dropped to zero and (2) time *T* amounts to approximately 9 h. This suggests that an approximately 2.5-fold increase in the ASC content in cell is already sufficient for the cell to follow a neural differentiation pattern.

The above data were obtained for the CRC operation in the norm. However, the model allows the relative contributions of CRC genes to its operation to be assessed as well by taking into account a mutation in each gene.

Modeling of CRC operation in the parental cell of mechanoreceptor in the presence of mutations in *AS-C* genes

As is known from experimental data, the mutations in *achaete-scute* genes appear as the absence of part of mechanoreceptors and, in several cases, even all mechanoreceptors of the standard set (Agol, 1931; Dubinin, 1932; Cabrera et al., 1994; Roark et al., 1995; Pi et al., 2001; Escudero et al., 2005; Acar et al., 2006; Usui et al., 2008; Garcıa-Bellido, de Celis, 2009).

Several numerical experiments were performed to assess the effects of mutations in *AS-C* genes on the CRC operation. The following parameters of system (1) were used in these experiments:

$D = 1.6$; $G = 1$; $m_x = 0.3$; $U = 1.1$; $S = 5.5$;
$a_1 = 2.9$; $n_1 = 1$; $b_1 = 1$; $a_4 = 5.8$; $n_4 = 1$; $b_4 = 5.6$;
$a_6 = 6$; $n_6 = 1$; $b_6 = 5.7$;
$C_y = 14.1$; $d_1 = 4.1$; $d_2 = 4.7$; $m_y = 0.5$;
$C_e = 2.9$; $d_3 = 7.5$; $m_e = 0.4$;
$\alpha_4 = 3$; $v_4 = 1.9$; $\beta_4 = 1.2$; $m_z = 1.6$;
$\alpha_5 = 14.8$; $v_5 = 1.1$; $\beta_5 = 14.8$; $m_u = 2.3$;
$\alpha_6 = 2$; $v_6 = 1$; $\beta_6 = 1$; $m_w = 1$;
$\alpha_7 = 4.5$; $v_7 = 3.1$; $\beta_7 = 0.5$; $m_p = 0.6$; $L = 1.1$;
$y_0 = 1.6$; $E_0 = 1.1$; $z_0 = 0.4$; $u_0 = 0$; $w_0 = 0$; $p_0 = 0$;
$T = 28$; $\tau = 12$.

In all equations of system (1) except for the first one, coefficients $k = 1$.

Table 1. Values of parameters k_{xi} and x_{0i} in modeling the effect of mutations in ASC on the content of the corresponding proteins in presumptive SOPC

Parameters	Number of experiment								
	1 (norm)	2	3	4	5	6	7	8	9
k_{xi}	1	0.9	0.6	0.5	0.4	0.3	0.2	0.1	0
x_{0i}	0.8	0.72	0.48	0.4	0.32	0.24	0.16	0.08	0

Table 1 lists the values of k_{xi} and x_{0i}. The value of parameter k_{xi} varies from 0 (complete absence of protein) to 1 (protein content in the norm) and from a biological standpoint, reflects the degree of influence of a mutation in *AS-C* on the content of ASC proteins. The smaller the value of k_{xi}, the lower is the content of the protein in the cell. Parameter x_{0i} defines the initial content of ASC proteins. In the numerical experiments, x_{01} is assumed to be 0.8, which corresponds to the norm, $k_{x1} = 1$ (see Fig. 2). Coefficients k_{xi} define a proportional decrease in the contents of proteins x_{0i} according to equation $x_{0i} = x_{01} \cdot k_{xi}$.

Figure 5 shows the results of numerical experiments. The above-described data demonstrate that the determination of a cell as an SOPC in the absence of mutations in the CRC genes becomes possible when the ASC content increases at least 2.5-fold as compared with the initial value (see Fig. 4). Thus, it is possible to assess the minimum k_{xi} value when this condition is met. The range of the content of ASC proteins permitting the determination of SOPC is colored turquoise. The plots showing the content of ASC proteins corresponds to the k_{xi} values at which the possibility of cell determination as an SOPC is retained.

The necessary level of the content of ASC proteins is achieved at $k_{xi} \geq 0.6$. The value of $k_{x3} = 0.6$ corresponds to a decrease in the content by 40 % relative to the initial values of the norm. From a biological standpoint, this means that a decrease in the ASC content in the cell by >40 % prohibits its differentiation according to a neural pathway and, consequently, entails the absence of mechanoreceptor.

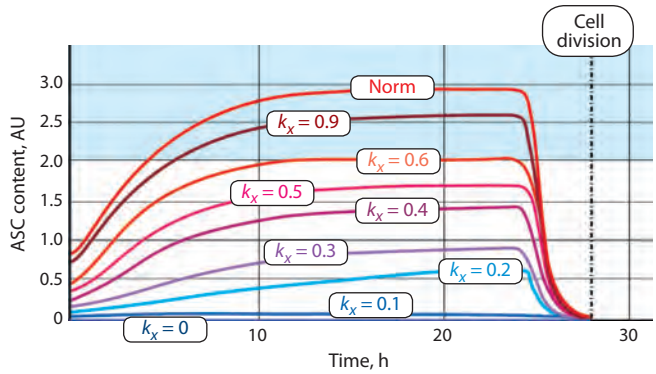


Fig. 5. Dynamics of the content of ASC proteins in the mechanoreceptor parental cell in the presence of mutations in the *achaete-scute* gene complex.

The region of ASC content at which SOPC determination is possible is colored turquoise.

Modeling of CRC operation in the presumptive SOPC in the presence of mutations in constituent genes

The CRC components are united via the intracellular system of positive and negative feedbacks (see Fig. 1), which strictly regulates the production and degradation of ASC proteins. Correspondingly, the mutations in each gene must influence the content of the corresponding proteins in the cell and have a certain phenotypic effect. Indeed, experiments have shown that the mutations of CRC genes appear as variations in the canonical architecture of bristle pattern, namely, changes in the number and/or positions of mechanoreceptors. The considered model that takes into account the mutational changes in CRC genes allows the degree and character of their effects on the dynamics of ASC content to be assessed. In the numerical experiments, coefficients k_y (for *hairy*), k_e (for *emc*), k_z (for *sens*), k_u (for *scrt*), k_w (for *chn*), and k_p (for *phyl*) were assumed to be zero, which corresponds to a complete absence of the corresponding proteins.

Several parameters remained constant:

$D = 1.6$; $G = 1$; $m_x = 0.3$; $U = 1.1$; $S = 5.5$;
$a_1 = 2.9$; $n_1 = 1$; $b_1 = 1$; $a_4 = 5.8$; $n_4 = 1$; $b_4 = 5.6$;
$a_6 = 6$; $n_6 = 1$; $b_6 = 5.7$;
$C_y = 14.1$; $d_1 = 4.1$; $d_2 = 4.7$; $m_y = 0.5$;
$C_e = 2.9$; $d_3 = 7.5$; $m_e = 0.4$;
$\alpha_4 = 3$; $v_4 = 1.9$; $\beta_4 = 1.2$; $m_z = 1.6$;
$\alpha_5 = 14.8$; $v_5 = 1.1$; $\beta_5 = 14.8$; $m_u = 2.3$;
$\alpha_6 = 2$; $v_6 = 1$; $\beta_6 = 1$; $m_w = 1$;
$\alpha_7 = 4.5$; $v_7 = 3.1$; $\beta_7 = 0.5$; $m_p = 0.6$; $L = 1.1$;
$T = 28$; $\tau = 12$;
$k_x = 1$; $x_0 = 0.8$.

The changing parameters are listed in Table 2: the k values of 0 or 1 mean the presence or absence of a mutation in a gene and parameters y_0, z_0, u_0, w_0, p_0, and E_0 specify the initial contents of the proteins Hairy, SENS, SCRT, CHN, PHYL, and EMC, respectively.

Figure 6 shows the results of numerical experiments. A comparison of the shapes of the plots shown in Figure 6 reveals a certain hierarchy of the CRC genes in their effects on the content of ASC proteins. This is reflected in the range of deviations from the plot that characterizes the dynamics of these proteins in the norm (in the absence of any mutations in all genes of the CRC). The larger the deviation, the stronger is the effect of an individual gene.

The *emc* (*emc*$^-$) and *hairy* (*hairy*$^-$) genes display the strongest effects because the mutations in these genes cause

Table 2. Values of changing parameters in modeling the effect of mutations in CRC genes on the content of ASC proteins

Mutation in gene	k_y	k_e	k_z	k_u	k_w	k_p	y_0	E_0	z_0	u_0	w_0	p_0
hairy$^-$	0	1	1	1	1	1	0	1.1	0.4	0	0	0
emc$^-$	1	0	1	1	1	1	1.6	0	0.4	0	0	0
sens$^-$	1	1	0	1	1	1	1.6	1.1	0	0	0	0
scrt$^-$	1	1	1	0	1	1	1.6	1.1	0.4	0	0	0
chn$^-$	1	1	1	1	0	1	1.6	1.1	0.4	0	0	0
phyl$^-$	1	1	1	1	1	0	1.6	1.1	0.4	0	0	0

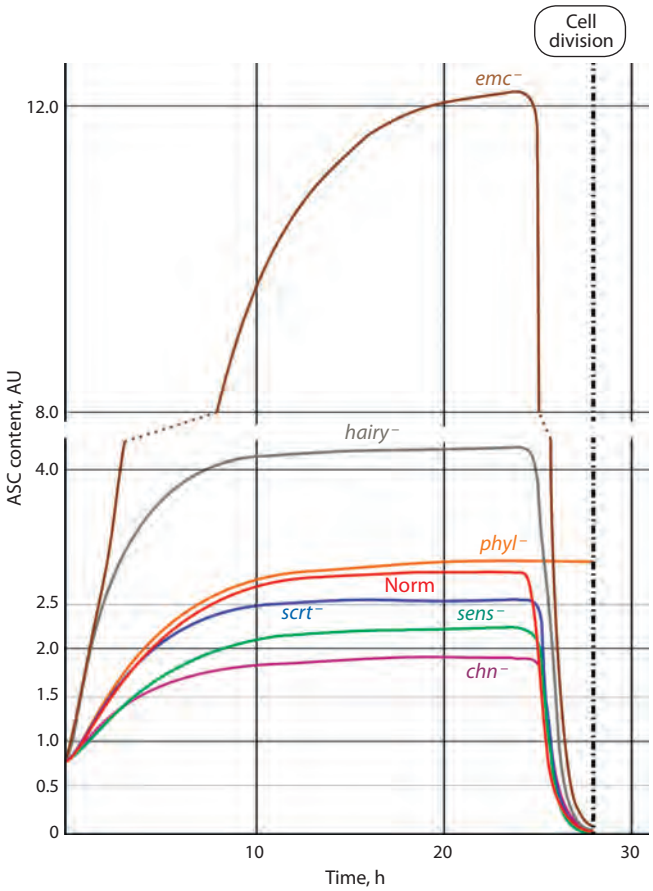


Fig. 6. Dynamics of the content of ASC proteins in the mechanoreceptor presumptive parent cell in the presence of mutations in CRC genes.

a considerable upward deviation of the ASC level from the normal characteristics. This is a biologically justified result since the EMC and Hairy proteins repress *AS-C* (Moscoso del Prado, Garcia-Bellido, 1984) so that the removal of this repression must appear as an increase in ASC content. A phenotypic manifestation of mutations consists in the development of additional mechanoreceptors (Ingham et al., 1985; de Celis et al., 1991). Presumably, a concurrent sharp and rapid increase in the ASC content in the cells of proneural cluster causes mistuning of intercellular interactions mediated by signaling pathways and the formation of several SOPCs in the proneural cluster rather than a single SOPC as in the norm.

The mutation in *chn* (*chn*$^-$) appears as a noticeable decrease in the ASC level (the corresponding curve lies below the curve for the norm). The effect is associated with the fact that the mutation in this gene causes the absence of the corresponding protein, which directly activates the *AS-C* genes and represses the *emc* and *hairy* genes (Escudero et al., 2005; Yamasaki et al., 2011). Correspondingly, the production of ASC proteins cannot reach the required values.

The mutations in genes *sens* (*sens*$^-$) and *scrt* (*scrt*$^-$) cause a less pronounced increase in the level of proteins, which also agrees with the known data on the functions of these genes in the CRC system and the manifestations of mutations in these genes. The SENS protein is known as a coactivator of *AS-C* activity and, consequently, the mutation will somewhat decrease the ASC production. The SCRT protein represses the *hairy* gene, thereby potentially increasing the ASC level, which, nonetheless, fails to reach the normal values because of the effects of other direct repressors of *AS-C* gene activity (Roark et al., 1995; Nolo et al., 2000) (see Fig. 1).

In the case of a mutation in the *phyl* gene (*phyl*$^-$), the ASC level expectedly remains on the reached plateau because the PHYL protein, responsible for its degradation, is not produced in this case (Chang et al., 2008). Thus, SOPC cannot transit to division and the phenotypic effect must appear as the absence of mechanoreceptor at its regular position. This conclusion is confirmed by experimental data (Pi et al., 2001).

Conclusion

The decades of the research into the system underlying the formation of bristle pattern on the head and body of drosophila have yielded a tremendous array of data giving the insight into individual mechanisms forming the basis for the function of this system. However, particular details of the morphogenesis of mechanoreceptor are still rather vague.

We have earlier demonstrated that the development of an individual mechanoreceptor and the overall bristle pattern are controlled by the central regulatory circuit, which determines the expression of *AS-C* genes and production of the corresponding proteins in the parental cell. A mathematical model of the CRC operation was elaborated taking into account all identified CRC components and the relations between them. This model allowed us to advance from a purely qualitative

description of the system controlling the content of ASC proteins and to succeed in clarification of its certain quantitative characteristics unknown earlier.

In particular, our numerical experiments suggest that the cell is determined as an SOPC when the ASC content increases approximately 2.5-fold relative to the initial level in the cells of proneural cluster. Individual elements of the circuit have different effects on the content of ASC proteins in the presumptive cell of mechanoreceptor. *AS-C*, the key CRC component, and the mutations that decrease the ASC content by more than 40 % have the most significant effect and cause the prohibition of SOPC segregation. As for the mutations in the remaining genes of the circuit, they change the level of ASC proteins to different degrees, with the most pronounced effects of mutations in the *emc* and *hairy* genes.

Thus, the model demonstrates that the CRC as a system is sensitive to changes in internal interactions and its robust operation, providing a certain dynamics of the level of ASC proteins, requires a concerted work of all components constituting the regulatory circuit. The model predictions are appropriate for experimental verification.

References

Acar M., Jafar-Nejad H., Giagtzoglou N., Yallampalli S., David G., He Y., Delidakis C., Bellen H.J. Senseless physically interacts with proneural proteins and functions as a transcriptional co-activator. *Development*. 2006;133(10):1979-1989. DOI 10.1242/dev.02372

Agol I.J. Step allelomorphism in *D. melanogaster*. *Genetics*. 1931; 16(3):254-266. DOI 10.1093/genetics/16.3.254

Audibert A., Simon F., Gho M. Cell cycle diversity involves differential regulation of Cyclin E activity in the *Drosophila* bristle cell lineage. *Development*. 2005;132(10):2287-2297. DOI 10.1242/dev.01797

Ayeni J.O., Audibert A., Fichelson P., Srayko M., Gho M., Campbell S.D. G2 phase arrest prevents bristle progenitor self-renewal and synchronizes cell division with cell fate differentiation. *Development*. 2016;143(7):1160-1169. DOI 10.1242/dev.134270

Bukharina T.A., Akinshin A.A., Golubyatnikov V.P., Furman D.P. Mathematical and numerical models of the central regulatory circuit of the morphogenesis system of *Drosophila*. *J. Appl. Ind. Math.* 2020;14(2):249-255. DOI 10.1134/S1990478920020040

Cabrera C.V., Alonso M.C. Transcriptional activation by heterodimers of the *achaete-scute* and *daughterless* gene products of *Drosophila*. *EMBO J*. 1991;10(10):2965-2973. DOI 10.1002/j.1460-2075.1991.tb07847.x

Cabrera C.V., Alonso M.C., Huikeshoven H. Regulation of *scute* function by *extramacrochaete in vitro* and *in vivo*. *Development*. 1994; 120(12):3595-3603. DOI 10.1242/dev.120.12.3595

Chang P.J., Hsiao Y.L., Tien A.C., Li Y.C., Pi H. Negative-feedback regulation of proneural proteins controls the timing of neural precursor division. *Development*. 2008;135(18):3021-3030. DOI 10.1242/dev.021923

Chasman D., Fotuhi Siahpirani A., Roy S. Network-based approaches for analysis of complex biological systems. *Curr. Opin. Biotechnol*. 2016;39:157-166. DOI 10.1016/j.copbio.2016.04.007

Corson F., Couturier L., Rouault H., Mazouni K., Schweisguth F. Self-organized Notch dynamics generate stereotyped sensory organ patterns in *Drosophila*. *Science*. 2017;356(6337):eaai7407. DOI 10.1126/science.aai7407

Cubas P., de Celis J.F., Campuzano S., Modolell J. Proneural clusters of *achaete-scute* expression and the generation of sensory organs in the *Drosophila* imaginal wing disc. *Genes Dev*. 1991;5(6):996-1008. DOI 10.1101/gad.5.6.996

de Celis J.F., Marí-Beffa M., García-Bellido A. Function of trans-acting genes of the *achaete-scute* complex in sensory organ patterning in the mesonotum of *Drosophila*. *Rouxs Arch. Dev. Biol*. 1991;200(2): 64-76. DOI 10.1007/BF00637186

Dubinin N.P. Step-allelomorphism in *D. melanogaster*. The allelomorphs *achaete2-scute10, achaete1-scute11* and *achaete3-scute13*. *J. Genet*. 1932;25(2):163-181. DOI 10.1007/BF02983250

Emmert-Streib F., Glazko G.V. Network biology: a direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med*. 2011;3(4):379-391. DOI 10.1002/wsbm.134

Escudero L.M., Caminero E., Schulze K.L., Bellen H.J., Modolell J. Charlatan, a Zn-finger transcription factor, establishes a novel level of regulation of the proneural *achaete/scute* genes of *Drosophila*. *Development*. 2005;132(6):1211-1222. DOI 10.1242/dev.01691

Furman D.P., Bukharina T.A. Genetic regulation of morphogenesis of *Drosophila melanogaster* mechanoreceptors. *Russ. J. Dev. Biol*. 2022;53(4):239-251. DOI 10.1134/S1062360422040038

Garcıa-Bellido A., de Celis J.F. The complex tale of the *achaete-scute* complex: a paradigmatic case in the analysis of gene organization and function during development. *Genetics*. 2009;182(3):631-639. DOI 10.1534/genetics.109.104083

Ghysen A., Thomas R. The formation of sense organs in *Drosophila*: a logical approach. *Bioessays*. 2003;25(8):802-807. DOI 10.1002/bies.10311

Giri R., Brady S., Papadopoulos D.K., Carthew R.W. Single-cell Senseless protein analysis reveals metastable states during the transition to a sensory organ fate. *iScience*. 2022;25(10):105097. DOI 10.1016/j.isci.2022.105097

Golubyatnikov V.P., Bukharina T.A., Furman D.P. A model study of the morphogenesis of *D. melanogaster* mechanoreceptors: the central regulatory circuit. *J. Bioinform. Comput. Biol*. 2015;13(1):1540006. DOI 10.1142/S0219720015400065

Hsu C.P., Lee P.H., Chang C.W., Lee C.T. Constructing quantitative models from qualitative mutant phenotypes: preferences in selecting sensory organ precursors. *Bioinformatics*. 2006;22(11):1375-1382. DOI 10.1093/bioinformatics/btl082

Huang F., Dambly-Chaudiere C., Ghysen A. The emergence of sense organs in the wing disc of *Drosophila*. *Development*. 1991;111(4): 1087-1095. DOI 10.1242/dev.111.4.1087

Ingham P.W., Pinchin S.M., Howard K.R., Ish-Horowicz D. Genetic analysis of the hairy locus in *Drosophila melanogaster*. *Genetics*. 1985;111(3):463-486. DOI 10.1093/genetics/111.3.463

Kawamori A., Shimaji K., Yamaguchi M. Temporal and spatial pattern of *dref* expression during *Drosophila* bristle development. *Cell Struct. Funct*. 2013;38(2):169-181. DOI 10.1247/csf.13004

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Y.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2): 833-850 (in Russian)

Marnellos G., Mjolsness E. A gene network approach to modeling early neurogenesis in *Drosophila*. In: Pacific Symposium on Biocomputing '98, January 4–9, 1998, in Hawaii. World Scientific Pub Co Inc., 1998;30-41

Meir E., von Dassow G., Munro E., Odell G.M. Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Curr. Biol*. 2002;12(10):778-786. DOI 10.1016/s0960-9822(02)00839-4

Moscoso del Prado J., Garcia-Bellido A. Genetic regulation of the *achaete-scute* complex of *Drosophila melanogaster*. *Wilehm Roux Arch. Dev. Biol*. 1984;193(4):242-245. DOI 10.1007/BF01260345

Nolo R., Abbott L.A., Bellen H.J. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila*. *Cell*. 2000;102(3):349-362. DOI 10.1016/s0092-8674(00)00040-4

Pi H., Wu H.J., Chien C.T. A dual function of *phyllopod* in *Drosophila* external sensory organ development: cell fate specification of sen-

sory organ precursor and its progeny. *Development*. 2001;128(14): 2699-2710. DOI 10.1242/dev.128.14.2699

Reeves N., Posakony J.W. Genetic programs activated by proneural proteins in the developing *Drosophila* PNS. *Dev. Cell*. 2005;8(3): 413-425. DOI 10.1016/j.devcel.2005.01.020

Roark M., Sturtevant M.A., Emery J., Vaessin H., Grell E., Bier E. *scratch*, a pan-neural gene encoding a zinc finger protein related to *snail*, promotes neuronal development. *Genes Dev*. 1995;9(19): 2384-2398. DOI 10.1101/gad.9.19.2384

Schlitt T., Palin K., Rung J., Dietmann S., Lappe M., Ukkonen E., Brazma A. From gene networks to gene function. *Genome Res*. 2003;13(12):2568-2576. DOI 10.1101/gr.1111403

Skeath J.B., Carroll S.B. Regulation of *achaete-scute* gene expression and sensory organ pattern formation in the *Drosophila* wing. *Genes Dev*. 1991;5(6):984-995. DOI 10.1101/gad.5.6.984

Usui K., Kimura K.I. Sequential emergence of the evenly spaced microchaetes on the notum of *Drosophila*. *Rouxs Arch. Dev. Biol*. 1993; 203(3):151-158. DOI 10.1007/BF00365054

Usui K., Goldstone C., Gibert J.M., Simpson P. Redundant mechanisms mediate bristle patterning on the *Drosophila* thorax. *Proc. Natl. Acad. Sci. USA*. 2008;105(51):20112-20117. DOI 10.1073/pnas.0804282105

Vaessin H., Brand M., Jan L.Y., Jan Y.N. *daughterless* is essential for neuronal precursor differentiation but not for initiation of neuronal precursor formation in *Drosophila* embryo. *Development*. 1994;120(4):935-945. DOI 10.1242/dev.120.4.935

Van Doren M., Powell P.A., Pasternak D., Singson A., Posakony J.W. Spatial regulation of proneural gene activity: auto- and cross-activation of *achaete* is antagonized by *extramacrochaetae*. *Genes Dev*. 1992;6(12B):2592-2605. DOI 10.1101/gad.6.12b.2592

Van Doren M., Bailey A.M., Esnayra J., Ede K., Posakony J.W. Negative regulation of proneural gene activity: *hairy* is a direct transcriptional repressor of *achaete*. *Genes Dev*. 1994;8(22):2729-2749. DOI 10.1101/gad.8.22.2729

Yamasaki Y., Lim Y.M., Niwa N., Hayashi S., Tsuda L. Robust specification of sensory neurons by dual functions of charlatan, a *Drosophila* NRSF/REST-like repressor of *extramacrochaetae* and *hairy*. *Genes Cells*. 2011;16(8):896-909. DOI 10.1111/j.1365-2443.2011.01537.x

Yasugi T., Sato M. Mathematical modeling of Notch dynamics in *Drosophila* neural development. *Fly* (*Austin*). 2022;16(1):24-36. DOI 10.1080/19336934.2021.1953363

Zhu X., Gerstein M., Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev*. 2007;21(9):1010-1024. DOI 10.1101/gad.1528707

ORCID ID

T.A. Bukharina orcid.org/0000-0002-9011-4196
V.P. Golubyatnikov orcid.org/0000-0002-9758-3833

Acknowledgements. The authors are sincerely grateful to A.A. Akin'shin for his helpful advice and criticism.

Funding. The work was supported by budget projects FWNR-2022-0020 (Institute of Cytology and Genetics SB RAS for T.A.B. and D.P.F) and FWNF-2022-0009 (Institute of Mathematics SB RAS for V.P.G.).

Conflict of interest. The authors declare no conflict of interest.

Received July 18, 2023. Revised September 20, 2023. Accepted September 25, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):755-767
DOI 10.18699/VJGB-23-88

Original Russian text https://vavilovj-icg.ru/

Bifurcation analysis of multistability and hysteresis in a model of HIV infection

I.V. Mironov[1, 2], M.Yu. Khristichenko[1, 3], Yu.M. Nechepurenko[1, 3], D.S. Grebennikov[2, 3], G.A. Bocharov[2, 3]

[1] Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences, Moscow, Russia
[2] Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation, Moscow, Russia
[3] Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences, Moscow, Russia
gbocharov@gmail.com

Abstract. The infectious disease caused by human immunodeficiency virus type 1 (HIV-1) remains a serious threat to human health. The current approach to HIV-1 treatment is based on the use of highly active antiretroviral therapy, which has side effects and is costly. For clinical practice, it is highly important to create functional cures that can enhance immune control of viral growth and infection of target cells with a subsequent reduction in viral load and restoration of the immune status. HIV-1 control efforts with reliance on immunotherapy remain at a conceptual stage due to the complexity of a set of processes that regulate the dynamics of infection and immune response. For this reason, it is extremely important to use methods of mathematical modeling of HIV-1 infection dynamics for theoretical analysis of possibilities of reducing the viral load by affecting the immune system without the usage of antiviral therapy. The aim of our study is to examine the existence of bi-, multistability and hysteresis properties with a meaningful mathematical model of HIV-1 infection. The model describes the most important blocks of the processes of interaction between viruses and the human body, namely, the spread of infection in productively and latently infected cells, the appearance of viral mutants and the development of the T cell immune response. Furthermore, our analysis aims to study the possibilities of transferring the clinical pattern of the disease from a more severe state to a milder one. We analyze numerically the conditions for the existence of steady states of the mathematical model of HIV-1 infection for the numerical values of model parameters corresponding to phenotypically different variants of the infectious disease course. To this end, original computational methods of bifurcation analysis of mathematical models formulated with systems of ordinary differential equations and delay differential equations are used. The macrophage activation rate constant is considered as a bifurcation parameter. The regions in the model parameter space, in particular, for the rate of activation of innate immune cells (macrophages), in which the properties of bi-, multistability and hysteresis are expressed, have been identified, and the features characterizing transition kinetics between stable equilibrium states have been explored. Overall, the results of bifurcation analysis of the HIV-1 infection model form a theoretical basis for the development of combination immune-based therapeutic approaches to HIV-1 treatment. In particular, the results of the study of the HIV-1 infection model for parameter sets corresponding to different phenotypes of disease dynamics (typical, long-term non-progressing and rapidly progressing courses) indicate that an effective functional treatment (cure) of HIV-1-infected patients requires the development of a personalized approach that takes into account both the properties of the HIV-1 quasispecies population and the patient's immune status.
Key words: mathematical model; HIV infection; ordinary differential equations; bifurcation analysis; stationary solutions; bistability; multistability; hysteresis; optimal control.

For citation: Mironov I.V., Khristichenko M.Yu., Nechepurenko Yu.M., Grebennikov D.S., Bocharov G.A. Bifurcation analysis of multistability and hysteresis in a model of HIV infection. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):755-767. DOI 10.18699/VJGB-23-88

Бифуркационный анализ мультистабильности и гистерезиса в модели ВИЧ-инфекции

И.В. Миронов[1, 2], М.Ю. Христиченко[1, 3], Ю.М. Нечепуренко[1, 3], Д.С. Гребенников[2, 3], Г.А. Бочаров[2, 3]

[1] Институт прикладной математики им. М.В. Келдыша Российской академии наук, Москва, Россия
[2] Первый Московский государственный медицинский университет им. И.М. Сеченова Министерства здравоохранения Российской Федерации, Москва, Россия
[3] Институт вычислительной математики им. Г.И. Марчука Российской академии наук, Москва, Россия
gbocharov@gmail.com

Аннотация. Инфекционное заболевание, вызванное вирусами иммунодефицита человека первого типа (ВИЧ-1), остается серьезной угрозой здоровью людей. Существующий подход к лечению ВИЧ-1 основан на применении высокоактивной антиретровирусной терапии, имеющей побочные эффекты для здоровья и высокую стоимость. Для практической медицины актуальной является задача поиска методов функционального лечения, связанных с интен-

сификацией иммунного контроля размножения вирусов и заражения клеток-мишеней с последующим снижением уровня вирусной нагрузки и восстановления иммунного статуса. Исследования в области иммунотерапии ВИЧ-1 находятся на стадии концептуальной разработки в силу сложности совокупности процессов, регулирующих динамику инфекции и иммунного ответа. По этой причине чрезвычайно актуальным является использование методов математического моделирования динамики ВИЧ-1 инфекции для теоретического анализа возможностей снижения вирусной нагрузки путем воздействия на иммунную систему без применения антивирусной терапии. Целью исследования было изучение, во-первых, свойств би-, мультистабильности и гистерезиса на примере содержательной модели ВИЧ-1 инфекции, которая описывает важнейшие блоки процессов взаимодействия вирусов и организма человека, а именно: распространение инфекции в продуктивно и латентно зараженных клетках, появление мутантов и развитие Т-клеточного иммунного ответа, и, во-вторых, возможностей перевода клинической картины заболевания из более тяжелого состояния в более легкое. В данной работе проведен численный анализ условий существования стационарных решений математической модели ВИЧ-1 инфекции для наборов параметров, отвечающих фенотипически различным вариантам течения инфекционного заболевания. Для этого использованы разработанные авторами методы бифуркационного анализа моделей, представляющих собой системы обыкновенных дифференциальных уравнений и дифференциальных уравнений с запаздыванием. В качестве бифуркационного параметра рассматривается константа скорости активации макрофагов. Определены области в пространстве параметров модели, в частности, для скорости активации клеток врожденного иммунитета (макрофагов), при которых имеют место свойства би-, мультистабильности и гистерезиса, и исследованы особенности кинетики перехода между устойчивыми положениями равновесия. В целом результаты бифуркационного анализа модели ВИЧ-1 инфекции формируют теоретическую основу для разработки комбинированных иммунотерапевтических воздействий для лечения ВИЧ-1. Результаты проведенного исследования модели ВИЧ-1 инфекции для параметров процессов, отвечающих разным фенотипам динамики заболевания (типичное, длительно не прогрессирующее и быстро прогрессирующее), указывают на то, что для эффективного функционального лечения больных ВИЧ-инфекцией требуется развитие персонализированного подхода, учитывающего как свойства популяции квазивидов ВИЧ-1, так и иммунный статус пациента.
Ключевые слова: математическая модель; ВИЧ-инфекция; обыкновенные дифференциальные уравнения; бифуркационный анализ; стационарные решения; бистабильность; мультистабильность; гистерезис; оптимальное управление.

Introduction

Human infectious disease caused by human immunodeficiency virus type 1 (HIV-1) remains a serious threat to human health worldwide, with the number of infections and deaths from associated complications of the order of 1.5×10^6 and 0.65×10^6, respectively (Landovitz et al., 2023). The current approach to HIV-1 treatment involves the continued use of highly active antiretroviral therapies (Gandhi et al., 2023), which inhibit various stages of the intracellular viral reproduction cycle and thus reduce the viral load in the patient's body. However, this approach has significant adverse side effects, as well as high treatment costs and suffers from interruption of the drug regimen (Trickey et al., 2022). For this reason, the search for therapies (Rasmussen, Søgaard, 2018; Niessl et al., 2020), including those related to the activation of immune control of virus reproduction and infection of target cells, and physiological mechanisms for boosting cellular homeostasis, is an urgent task (Grossman et al., 2020) that needs to be addressed following a systems immunology approach (Ludewig et al., 2012, Villani et al., 2018). The research in the field of immunotherapy-based treatment of HIV-1 is at the conceptualization stage due to the complexity of the set of processes that regulate the dynamics of infection and immune response (Landovitz et al., 2023). In this regard, the use of methods of mathematical modeling of HIV-1 infection dynamics is a tool for theoretical analysis of opportunities for viral load reduction by influencing the immune system without the use of antiviral therapy (Bocharov et al., 2022).

As has been previously noted (Bocharov et al., 2021), one of the goals of the development of mathematical models created to describe and study the dynamics of infectious diseases is the analysis of the characteristics of the dynamics sensitivity to influences of different nature, for example, in relation to perturbations of the parameters of regulatory processes or the state of the system in phase space. The results of modeling allow one to translate into a rational mode the design of combined control actions for correction of unfavorable infection course, in particular, from the region characterized by a high viral load to the region with a low viral load. The feasibility of the corresponding transitions is determined by the fundamental characteristics of the modeled system – the presence of bistability and/or multistability and hysteresis. Bistability, as an ability of the system "virus–human host" to coexist in two stable steady states, justifies the search for functional cure regiments of viral infection leading to transition from a chronic stable steady state with a higher viral load to a more favorable stable steady state with a lower viral load by inducing the activation of immune system components. The presence of the hysteresis property in bifurcation curves of a dynamical system makes the backstory significant, in particular, the critical importance of the branch on which the steady state of the system has been located before the subsequent change of bifurcation parameters (Khristichenko et al., 2022).

Research on mathematical modeling of HIV-1 infection dynamics in the human host has been actively developing for the last 30 years (Perelson, Nelson, 1999; Nowak, May, 2000). The key research areas were systematically presented in our earlier review (Bocharov et al., 2012). The main focus of the related papers is aimed at studying the infection kinetics during the application of antiretroviral therapy using low-dimensional models (Akın et al., 2020). Models of HIV-1 infection that consider the development of antiviral immune response are also related to the problem of estimating the infection parameters from individual patient's data (Banks et al., 2017). Conceptual aspects of HIV-1 infection dynamics, such as multistability and hysteresis, remain an underexplored

problem and the study of steady states is mainly reduced to elucidating the conditions for the existence of an infection-free equilibrium and the state of the infected organism as a function of the model parameters combined together in the basic reproductive number (Perelson, Nelson, 1999; Nowak, May, 2000).

The aim of this study is to investigate, firstly, the properties of bi-, multistability and hysteresis for a model of HIV-1 infection that describes the most important blocks of virus–human host interaction processes for sets of model parameters corresponding to different phenotypes of disease dynamics, i. e. known as typical progression, long-term non-progression and rapid progression courses, and, secondly, the conditions for transferring the mode of disease course from a more severe to a less severe state.

The specific objectives of this research include the bifurcation analysis of the model of the HIV-1 infection to identify the ranges of parameter values in which several steady states coexist, and the study of transitions between them, which are characterized by dependence on the prehistory of the state of the "virus–human host" system (hysteresis property). As a reference mathematical model for the study of stationary modes of HIV-1 infection dynamics and transitions between them, we consider a previously developed mathematical model (Hadjiandreou et al., 2009), which is characterized by the following essential properties:

- it describes the entire kinetics of infectious disease from early infection to the AIDS stage,
- it comprises a fairly complete spectrum of infection and immune response processes,
- the model parameters corresponding to different phenotypes of infection dynamics are provided,
- the description of antiretroviral therapy is included,
- the antiretroviral therapy with consideration of side effects is discussed and analyzed as an optimal control problem.

Previously, we used this model to develop a more complete description of the immune response to HIV infection that takes into account neuroendocrine regulation of the immune system, in particular, the influence of hormones (TSH, T3, T4) on the immune response, and to examine an optimal antiviral therapy on its basis (Savinkova et al., 2019).

The present work consists of four sections. Section "Materials and metods" describes the considered mathematical model of HIV-1 infection and the numerical methods used to analyze the model. Section "Results" presents the results of studying the steady states of the model system by tracing them by varying the model parameters, and the analysis of steady state changes under therapeutic interventions, which are described in the model as additional control variables on the right-hand sides of the model equations, i. e. in the terms for infection of target cells and virus replication processes. The application of the results of this work to the theoretical development of new approaches to HIV-1 treatment is discussed in Section "Discussion".

Materials and methods

Let us define the basic concepts that will be used throughout the paper.

- "Functional cure of HIV-1 infection" is an approach to therapy of the chronic infection associated with activation of immune control of viral replication and target cell infection that allows to exclude the use of antiretroviral drugs.
- "Bi-(multi)stability" is the property of a dynamical system to have two (or more) stable steady state solutions at the same parameter values.
- "Hysteresis" is a property of a dynamical system that is characterized by the dependence of its steady state on the backstory curve for the parameter being varied, which can be used for transition from one steady state to another by varying the parameters.

Mathematical model of HIV infection

The considered mathematical model of HIV infection is formulated in (Hadjiandreou et al., 2009) as a system of 11 ordinary differential equations. It describes the rate of change in time of the following concentrations: wild-type (wt) virus V_1; mutated virus V_2; CD4$^+$ T cells T; wt virus-infected CD4$^+$ T cells, T_1; CD4$^+$ T cells infected with mutated T_2; latently wt virus infected T cells T_{L1}; CD4$^+$ T cells, latently-infected with mutated virus-infected T cells T_{L2}; macrophages M; wt virus-infected macrophages M_1; macrophages infected with mutated virus M_2; cytotoxic CD8$^+$ T lymphocytes CTL. The system includes three blocks of equations: (1) the CD4$^+$ T cell block, (2) the macrophage and CTL block, and (3) the wild-type and mutant virus block.

The first block includes the equation for CD4$^+$ T cells:

$$\frac{dT}{dt} = s_1 + \frac{p_1(V_1+V_2)T}{V_1+V_2+S_1} - (1-u_1)(k_1V_1+k_2M_1)T - \\ - \varphi(k_1V_2+k_2M_2)T + rT\left(1-\frac{T+T_1+T_2+T_{L1}+T_{L2}}{T_{\max}}\right) - \delta_1 T, \quad (1)$$

where the 1st term describes the constant influx of CD4$^+$ T cells from the thymus, the 2nd term describes antigen-induced division, the 3rd term describes the loss due to infection by wt viruses and population of wt virus-infected macrophages, the 4th term describes the infection by mutated viruses and population of mutant virus-infected macrophages, the 5th term describes the homeostatic proliferation, and the 6th term describes natural cell death. It also includes the following two equations for infected CD4$^+$ T cells:

$$\frac{dT_1}{dt} = (1-u_1)\psi(k_1V_1+k_2M_1)T + \alpha_1 T_{L1} - \delta_2 T_1 - k_3 T_1 CTL \quad (2)$$

and

$$\frac{dT_2}{dt} = \psi\varphi\,(k_1V_2+k_2M_2)T + \alpha_1 T_{L2} - \delta_2 T_2 - k_3 T_2 CTL, \quad (3)$$

where in each equation, the 1st term describes population growth due to infections by wt or mutated virus and wt and mutated virus-infected macrophages; the 2nd term describes the transition of latently infected cells to productively infected cells; the 3rd term describes natural cell death, and the 4th term describes the CTL-mediated destruction of infected cells. The last two equations of the first block read as follows:

$$\frac{dT_{L1}}{dt} = (1-u_1)(1-\psi)(k_1V_1+k_2M_1)T - \alpha_1 T_{L1} - \delta_3 T_{L1} \quad (4)$$

and

$$\frac{dT_{L2}}{dt} = (1-\psi)\varphi(k_1V_2+k_2M_2)T - \alpha_1 T_{L2} - \delta_3 T_{L2}, \quad (5)$$

where in each of the equations the 1st term describes population growth due to infection by wt or mutated viruses and wt or mutated virus-infected macrophages; the 2nd term describes

the transition of latently infected cells to productively infected cells, and the 3rd term describes natural cell death.

The second block for macrophage and *CTL* dynamics consists of the equation:

$$\frac{dM}{dt} = s_2 + \frac{p_2(V_1+V_2)M}{V_1+V_2+S_2} - (1-f_1u_1)k_4V_1M - \varphi k_4V_2M - \delta_4 M, \quad (6)$$

where the 1st term describes the constant influx of cells from the bone marrow, the 2nd term describes the process of activation of macrophages with the possibility of their subsequent division due to chronic inflammation caused by HIV-1 infection, the 3rd term describes the infection of macrophages by wt viruses, the 4th term describes infection of macrophages by mutated viruses, and the 5th term describes natural death. This block also includes two equations for infected macrophages:

$$\frac{dM_1}{dt} = (1-f_1u_1)k_4V_1M - \delta_5 M_1 - k_5M_1CTL \quad (7)$$

and

$$\frac{dM_2}{dt} = \varphi k_4V_2M - \delta_5 M_2 - k_5M_2CTL, \quad (8)$$

where the 1st term describes the population growth due to infection of macrophages by wt or mutated viruses, the 2nd term describes natural death, and the 3rd term describes destruction by *CTL* effect. Finally, it includes the equation:

$$\frac{dCTL}{dt} = s_3 + k_6(T_1+T_2)CTL + k_7(M_1+M_2)CTL - \delta_6 CTL, \quad (9)$$

where the 1st term describes a constant influx of $CD8^+$ T cells from the thymus, the 2nd term describes the clonal proliferation induced by infected $CD4^+$ T cells, the 3rd term describes the clonal proliferation induced by infected macrophages, and the 4th term describes cell death.

The third block of wt and mutant virus dynamics consists of two equations

$$\frac{dV_1}{dt} = (1-u_2)(1-\mu)k_8T_1 + (1-f_2u_2)(1-\mu)k_9M_1 + \\ + \mu\varphi k_8T_2 + \mu\varphi k_9M_2 - (k_{10}T+k_{11}M)V_1 - k_{12}V_1M - \delta_7V_1 \quad (10)$$

and

$$\frac{dV_2}{dt} = (1-\mu)\varphi k_8T_2 + (1-\mu)\varphi k_9M_2 + (1-u_2)\mu k_8T_1 + \\ + (1-f_2u_2)\mu k_9M_1 - (k_{10}T+k_{11}M)V_2 - k_{12}V_2M - \delta_7V_2, \quad (11)$$

where in each of the equations the 1st term describes virus production by infected $CD4^+$ T cells, the 2nd term describes virus production by infected macrophages, the 3rd term describes virus production by infected $CD4^+$ T cells following mutations, the 4th term describes virus production by infected macrophages following mutations, the 5th term describes virus uptake by cells when infecting target cells, the 6th term describes virus elimination by the innate system immune cells, and the 7th term describes natural virus death. The biological meaning of the system parameters and their acceptable ranges are taken from the original work (Hadjiandreou et al., 2009) and summarized in Table 1.

Optimal control problem

In the article (Hadjiandreou et al., 2009), the possibility of optimizing the mode of administration of protease (RDV) and reverse transcriptase (3TC, ZDV) inhibitors was studied. Their concentrations are described by the following equations,

$$C_i(t) = C_i(t_l)e^{-k_e^i(t-t_l)} + \\ + \frac{F_iD_i}{V_c^i}\frac{k_a^i}{k_a^i+k_e^i}\left[e^{-k_e^i(t-t_l)} - e^{-k_a^i(t-t_l)}\right] \quad (i = 1, 2, 3), \quad (12)$$

where i is the drug index, t_l is the time of drug administration, D_i is the dose of the administered drug, F_i is the absolute bioavailability of the drug, k_a^i is the drug absorption rate, $k_e^i = Cl_i/V_c^i$ is the drug elimination rate constant (Cl_i is the elimination rate, V_c^i is the drug distribution volume). The values of all the above parameters are summarized in Table 2.

Control variables u_1 and u_2 were assumed to depend on the concentration of these drugs as follows:

$$u_1(t) = \frac{(C_2(t)/IC_{50}^2) + (C_3(t)/IC_{50}^3)}{1 + (C_2(t)/IC_{50}^2) + (C_3(t)/IC_{50}^3)},$$

$$u_2(t) = \frac{C_1(t)}{C_1(t) + \omega IC_{50}^1},$$

where $C_i(t)$ is the concentration of drug i in plasma at time t, IC_{50}^i is the average concentration of the drug that provides 50 % inhibition of virus replication processes. The parameter ω is a conversion factor between the value of the average concentration of the drug providing 50 % inhibition of virus replication processes IC_{50} obtained *in vitro*, and the same value obtained *in vivo*. The value ω = 1 was used in the computations. The goal of optimization in the original work was to achieve the maximum concentration of $CD4^+$ T cells (variable T in the system (1–11)) with the minimum index of adverse drug effects (Joly, Pinto, 2006)

$$S_e = \sum_{i=1}^{N} \overline{e}_i \frac{C_i(t)}{\overline{C}_i},$$

where $$\overline{e}_i = \frac{e_i}{\max\limits_i e_i}, \quad e_i = \sum_{j\in J_i} q_j h_{i,j}.$$

Here J_i is the set of side effects from the drug i, $\overline{C_i}$ is the average concentration of the drug i at steady state at standard dosage, that is, according to the regulation rules of antiretroviral therapy, e_i ($\overline{e_i}$) is the magnitude (normalized value) of the side effect caused by the drug i at the standard dosage, $h_{i,j}$ is the frequency of occurrence of the side effect j when exposed to the drug i at the standard dosage, and q_j is the relative magnitude of the side effect j, that is, its "undesirability".

The optimal control problem was formulated as a problem of maximizing the functional that depends on the concentration of $CD4^+$ T lymphocytes and the severity of side effects:

$$\int_{t_0}^{t_f} [A_1T - A_2S_e]dt \to \max_{C_1,C_2,C_3}, \quad T \ge T_{AIDS},\ t_0 \le t \le t_f,$$

where $A_1 = 1$ and $A_2 = 1000$ are weight coefficients, t_0 and t_f specify the optimization time interval, and the condition $T \ge T_{AIDS}$ prevents the cell concentration from falling below the threshold corresponding to the development of AIDS (200 cells/mm^{-3}).

Three sets of parameter values corresponding to different phenotypic variants of HIV infection course were considered:

Table 1. Biological meaning of the model parameters and their admissible ranges

Parameter	Biological meaning	Range
s_1	Rate constant for the influx of new uninfected CD4$^+$ T cells	5–36 mm^{-3}d^{-1}
s_2	Rate constant for the influx of new macrophages	0.03–0.015 mm^{-3}d^{-1}
s_3	Rate constant for the formation of new cytotoxic T lymphocytes	–
p_1	Activation rate constant for clonal expansion of CD4$^+$ T cells due to the immune response	0.01–5 d^{-1}
p_2	Rate constant of macrophage activation	–
S_1	Saturation constant	1–188 mm^{-3}
S_2	Saturation constant	–
k_1	Infection rate constant of CD4$^+$ T cells	10^{-8}–10^{-2} mm^{3}d^{-1}
k_2	Infection rate constant of CD4$^+$ T cells	10^{-6} mm^{3}d^{-1}
k_3	Rate constant of killing of infected CD4$^+$ T cells by cytotoxic T lymphocytes	10^{-4} – 1 mm^{3}d^{-1}
k_4	Rate constant of macrophage infection by viruses	$4.7\cdot10^{-9}$ – 10^{-3} mm^{3}d^{-1}
k_5	Rate constant of killing of infected macrophages by cytotoxic T lymphocytes	–
k_6	Proliferation rate constant of cytotoxic T lymphocytes stimulated by infected CD4$^+$ T cells	10^{-6}–10^{-3} mm^{3}d^{-1}
k_7	Proliferation rate constant of cytotoxic T lymphocytes stimulated by infected macrophages	–
k_8	Rate constant of virus production by infected CD4$^+$ T cells	$2.4\cdot10^{-1}$ – $5\cdot10^2$ d^{-1}
k_9	Rate constant of virus production by infected macrophages	$5\cdot10^{-3}$ – $3\cdot10^2$ d^{-1}
k_{10}	Rate constant of viral reduction rate associated with CD4$^+$ T cells infection expenditure	10^{-8} – 10^{-2} mm^{3}d^{-1}
k_{11}	Rate constant of virus loss for infection of macrophages	$4.7\cdot10^{-9}$ – 10^{-3} mm^{3}d^{-1}
k_{12}	Rate constant of virus elimination mediated by immune response	–
δ_1	Rate constant of natural death of uninfected CD4$^+$ T cells	0.01–0.02 d^{-1}
δ_2	Rate constant of natural death of infected CD4$^+$ T cells	0.24–0.7 d^{-1}
δ_3	Rate constant of natural death of latently infected CD4$^+$ T cells	0.02–0.069 d^{-1}
δ_4	Rate constant of natural death of macrophages	0.005 d^{-1}
δ_5	Rate constant of natural death of infected macrophages	0.005 d^{-1}
δ_6	Rate constant of natural death of cytotoxic T lymphocytes	0.015–0.05 d^{-1}
δ_7	Rate constant of natural virus death	2.39–13 d^{-1}
α_1	Activation constant of latently infected CD4$^+$ T cells	–
ψ	The fraction of CD4$^+$ T cells that become productively infected, and (1 – ψ) stand for the fraction which becomes latently infected	0.93–0.98
φ	Factor describing the reduction of the infection rate and replication of the mutated virus	0.1–0.9
r	Rate constant of the homeostatic proliferation of uninfected CD4$^+$ T cells	0.03 d^{-1}
T_{max}	Maximum concentration of CD4$^+$ T cells	1500–2000 mm^{-3}
μ	The fraction of viruses that mutate	$3\cdot10^{-5}$ –10^{-3}
f_i	The reduction of the treatment efficacy for macrophages as compared to CD4$^+$ T cells	0.34

Table 2. Parameter values for the pharmacokinetic equations (12)

Parameter	RDV, C_1	3TC, C_2	ZDV, C_3
D [mg]	600	150	300
k_a [d^{-1}]	2.4	12	12
Cl [$L\cdot$d^{-1}]	$1.48\cdot10^4$	$5.6\cdot10^2$	$2.69\cdot10^3$
V_c [L]	28.7	91	112
F	1.0	0.86	0.64
τ [d]	0.5	0.5	0.5
IC_{50} [mg$\cdot L^{-1}$]	0.11	0.34	0.13

Table 3. Values of model parameters (1–11) corresponding to a typical course of HIV infection (TP)

Parameter	Values	Parameter	Values	Parameter	Values
s_1	10 mm^{-3}d^{-1}	k_5	3·10^{-6} mm^{3}d^{-1}	δ_4	5·10^{-3} d^{-1}
s_2	0.15 mm^{-3}d^{-1}	k_6	3.3·10^{-4} mm^{3}d^{-1}	δ_5	5·10^{-3} d^{-1}
s_3	5 mm^{-3}d^{-1}	k_7	6·10^{-9} mm^{3}d^{-1}	δ_6	0.015 d^{-1}
p_1	0.16 d^{-1}	k_8	5.37·10^{-1} d^{-1}	δ_7	2.39 d^{-1}
p_2	0.15 d^{-1}	k_9	2.85·10^{-1} d^{-1}	α_1	3·10^{-4} d^{-1}
S_1	55.6 mm^{-3}	k_{10}	7.79·10^{-6} mm^{3}d^{-1}	ψ	0.97
S_2	188 mm^{-3}	k_{11}	10^{-6} mm^{3}d^{-1}	φ	0.9
k_1	3.87·10^{-3} mm^{3}d^{-1}	k_{12}	4·10^{-5} mm^{3}d^{-1}	r	0.03 d^{-1}
k_2	10^{-6} mm^{3}d^{-1}	δ_1	0.02 d^{-1}	T_{max}	1500 mm^{-3}
k_3	4.5·10^{-4} mm^{3}d^{-1}	δ_2	0.28 d^{-1}	μ	0.001
k_4	5.22·10^{-4} mm^{3}d^{-1}	δ_3	0.05 d^{-1}	f_i	0.34

Table 4. Values of model parameters (1–11) corresponding to different HIV infection phenotypes

Parameter	RP	TP	LTNP	Parameter	RP	TP	LTNP
p_1	0.13 d^{-1}	0.16 d^{-1}	0.20 d^{-1}	k_5	2.64·10^{-6} mm^{3}d^{-1}	3·10^{-6} mm^{3}d^{-1}	6.6·10^{-6} mm^{3}d^{-1}
p_2	0.1365 d^{-1}	0.15 d^{-1}	0.1638 d^{-1}	k_6	2.9·10^{-4} mm^{3}d^{-1}	3.3·10^{-4} mm^{3}d^{-1}	3.63·10^{-4} mm^{3}d^{-1}
S_1	50.0 mm^{-3}	55.6 mm^{-3}	55.6 mm^{-3}	k_7	5.28·10^{-9} mm^{3}d^{-1}	6·10^{-9} mm^{3}d^{-1}	6.6·10^{-9} mm^{3}d^{-1}
S_2	169.2 mm^{-3}	188 mm^{-3}	188 mm^{-3}	k_{12}	3.52·10^{-5} mm^{3}d^{-1}	4·10^{-5} mm^{3}d^{-1}	4.4·10^{-5} mm^{3}d^{-1}
k_3	3.96·10^{-4} mm^{3}d^{-1}	4.5·10^{-4} mm^{3}d^{-1}	9.9·10^{-4} mm^{3}d^{-1}	r	0.03	0.03	0.072

typical progression course (TP), rapid progression course (RP) and long-term non-progression course (LNTP). The parameter values in these sets are summarized in Tables 3 and 4.

In the original study (Hadjiandreou et al., 2009), a more effective regimen of drug administration based on optimization results was found to be superior to the standard treatment regimen for the parameters of a patient with a typical course of HIV infection with an initial CD4$^+$ T cell concentration equal to 350 mm^{-3}. While the standard treatment of the patient managed to keep the concentration of CD4$^+$ T cells above the AIDS threshold for about 2,500 days, the treatment regimen based on the optimization results extended it to longer than 10,000 days with a more than four times lower value of the side-effect index S_e.

Numerical methods

To numerically integrate the system (1–11), we used an implicit second-order BDF2 scheme (Hairer et al., 1987) on a sufficiently fine uniform grid built in half-interval $t \geq 0$. The accuracy of the results for the selected grid step was checked in all experiments requiring time integration. Symbolic computation methods (Geddes et al., 1992) implemented in the NSolve procedure of Mathematica were used to find steady states for given parameter values. To trace the solutions by varying parameters (i.e., to investigate the dependence of steady states of the system (1–11) on the parameters), we used the original algorithm proposed in (Nechepurenko et al., 2020). The study of asymptotic stability of a given steady state was reduced to the computation of eigenvalues of the system linearized with respect to this steady state and checking that all the found eigenvalues lie strictly in the left half-plane. To compute the eigenvalues, we used the standard QR algorithm (Golub, Van Loan, 1989).

Results

Bifurcation analysis

This section presents the results of the study of the dependence of steady states of the model of HIV infection dynamics on the activation rate of macrophages p_2 leading to their division, for three sets of values of the other parameters as given in "Materials and methods". Earlier, for the mathematical model of hepatitis B virus infection we showed the key role of the activation rate of innate immunity in the determination of different modes of hepatitis dynamics (Khristichenko et al., 2023), the analog of which in this model is p_2. The parameter p_2 was varied in the range from 0.13 to 0.17. The range of variation of the parameter p_2 was chosen to cover those values that correspond to the kinetics of innate immunity activation for three different modes of disease course (typical progression, long-term non-progression and rapid progression) shown in Table 4.

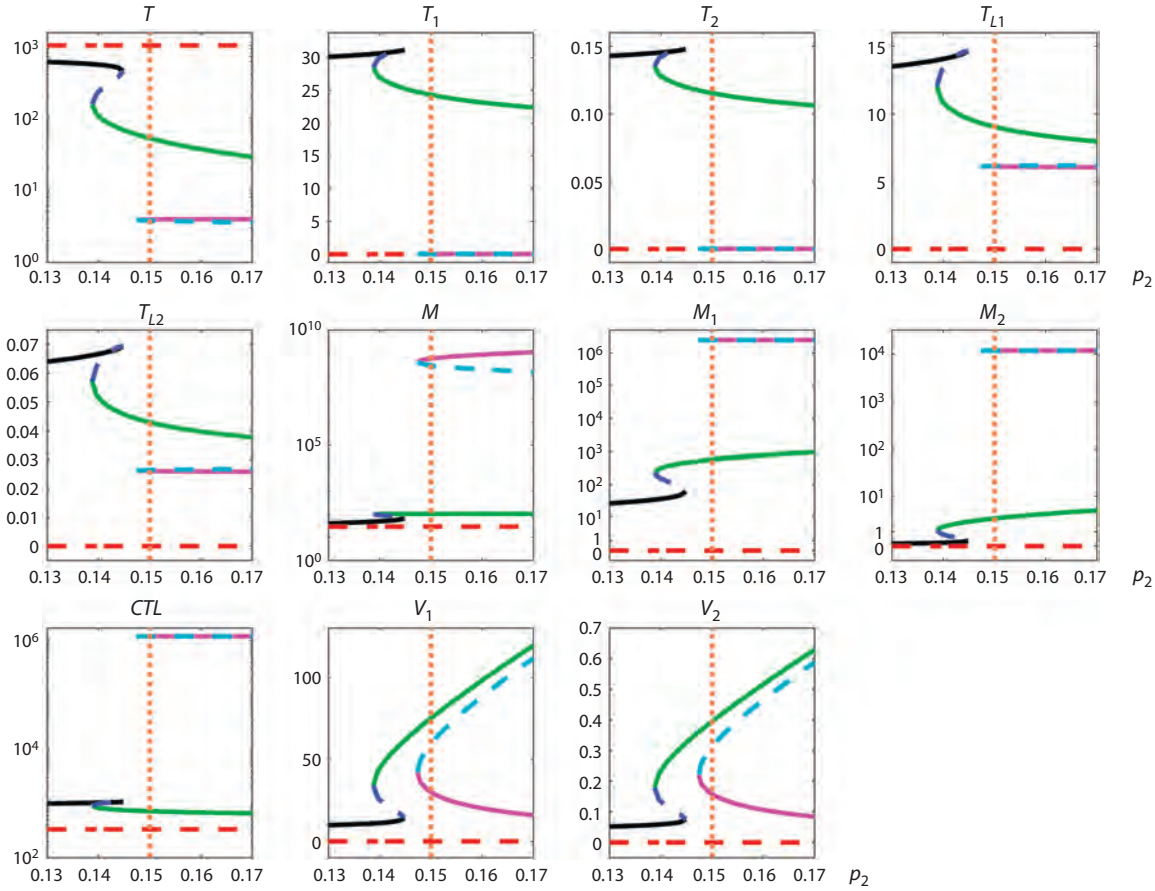


Fig. 1. Tracing of steady states by parameter p_2 for typical progression (TP) showing the presence of bistability and hysteresis.

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, and different colors indicate different steady states. The vertical orange dotted line indicates the value of the parameter p_2 corresponding to a TP course of infection.

Figures 1–3 summarize the tracing results. The vertical orange dotted line indicates the value of parameter p_2 taken from the corresponding parameter set, solid lines show stable steady states and dashed lines show unstable steady states, different colors indicate different steady states. It should be noted that the leading eigenvalues of the linearized equations corresponding to unstable steady states were real in all cases considered. Therefore, stable periodic solutions, which could otherwise be in the neighborhood of unstable steady states (Khristichenko, Nechepurenko, 2021), were absent in the considered cases.

Bistability. For a typical progression (TP) infection course (see Fig. 1), it can be seen that bistability is present at $0.138 < p_2 < 0.144$ (black and green lines) and at $0.147 < p_2 < 0.17$ (green and purple lines). For a rapid progression (RP) course (see Fig. 2), bistability is present at $0.135 < p_2 < 0.17$ (black and green lines). For a long-term non-progression (LTNP) course (see Fig. 3), bistability is present at $0.161 < p_2 < 0.17$ (blue and purple lines). The presence of two different stable steady states means that there is a possibility of establishment of a milder or more severe form of the disease in the same patient, depending on the patient's backstory. Note that for a RP infection course, both equilibria are characterized by a depleted $CD4^+$ T cell population, with macrophages being the dominant source of viruses. For such patients, the task of treatment becomes more complicated, because it is necessary to find changes in the system parameters, at which the equilibrium with a higher level of $CD4^+$ T cells would emerge.

In general, the obtained estimates of the areas of bistability together with the characteristics of bifurcation diagrams show that as the severity of the infection increases, i. e., as we move from long-term non-progressors to typical progressors and further to rapid progressors, the range of values of the activation rate of innate immunity cells, at which bistability takes place, increases. At the same time, some features of bifurcation diagrams change as well. These specific features of the response of an HIV-infected patient should be taken into account and used in the design of immunomodulatory regimes.

Multistability. The multistability property, as shown in Figure 3, occurs in the case of a LTNP infection course at $0.146 < p_2 < 0.161$ (black, blue and purple lines). The respective stable steady states correspond to different forms of the disease course in terms of the severity and efficacy of the immune response. Thus, the spectrum of possible stable steady-state modes of HIV-1 infection dynamics is more diverse in long-term non-progressors.

Hysteresis. The presence of the hysteresis property for this model is demonstrated in Figure 1. In particular, the behavior

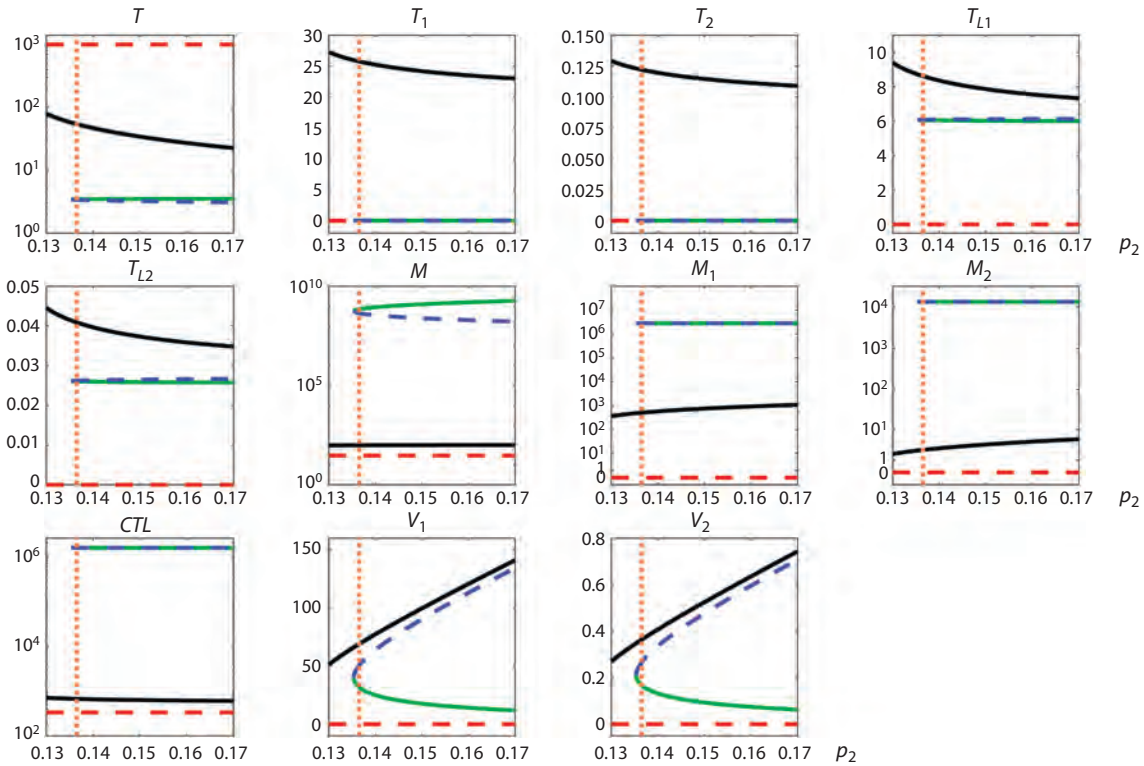


Fig. 2. Tracing of steady states by p_2 for rapid progression (RP) showing bistability.

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, and different colors indicate different steady states. The vertical orange dotted line indicates the value of the parameter p_2 corresponding to a RP course of the infection.

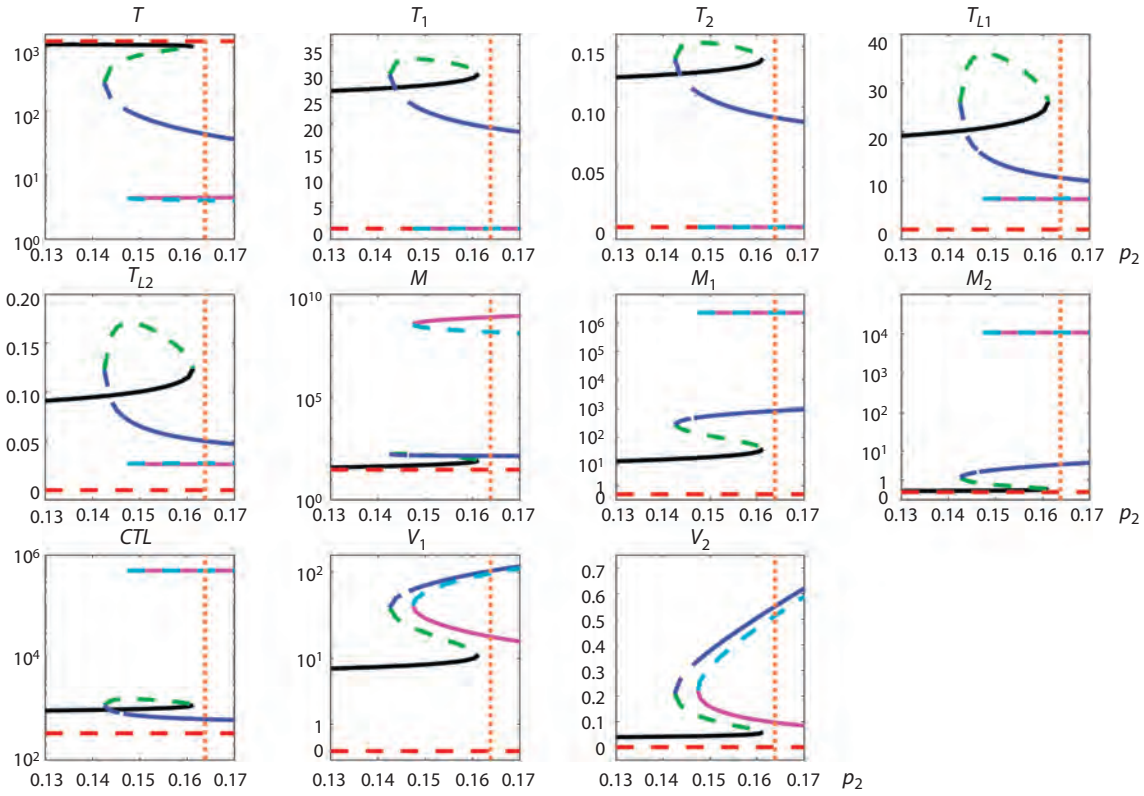


Fig. 3. Tracing of steady states by p_2 for long-term non-progression (LTNP) showing multistability.

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, and different colors indicate different steady states. The vertical orange dotted line indicates the value of the parameter p_2 corresponding to a LTNP course of the infection.

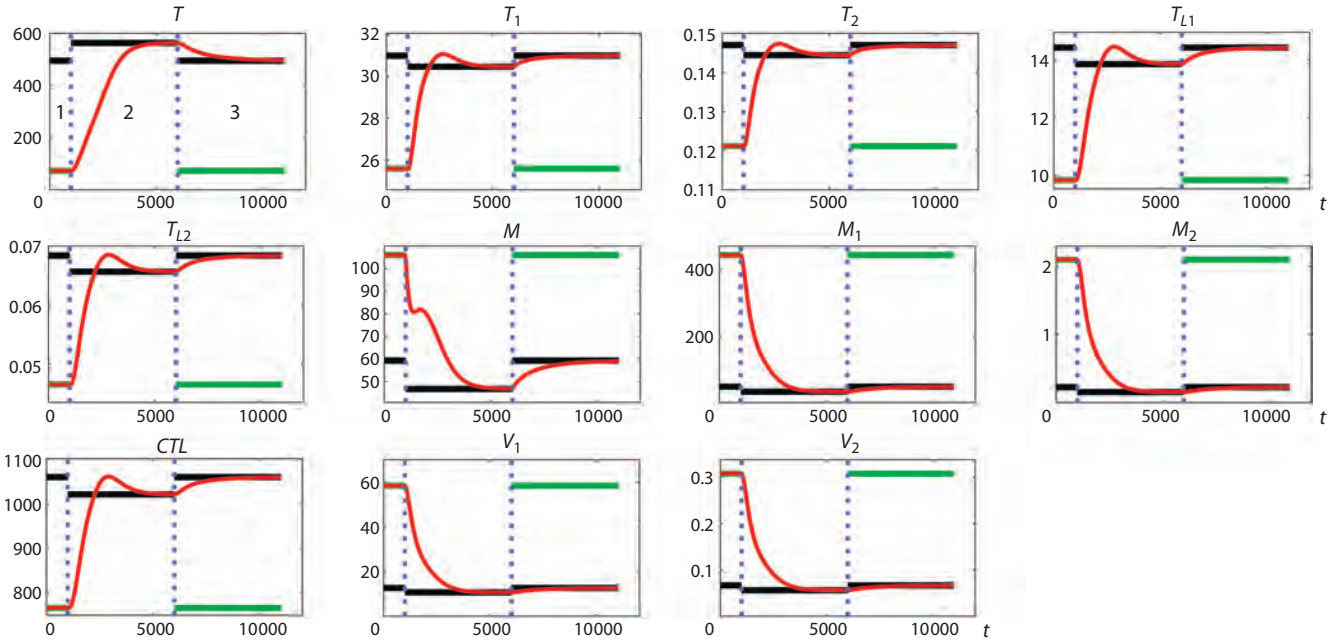


Fig. 4. Demonstration of the transition kinetics from a less favorable steady state to a more favorable steady state in the presence of hysteresis for typical progression (TP), where $p_2 = 0.143$ in regions 1 and 3, and $p_2 = 0.136$ in region 2.

The horizontal axis indicates time in days. The red solid line shows the dynamics of the model variables, the blue vertical dotted lines show the partitioning into regions 1–3, and the horizontal solid lines show the stable steady states of the variables in these regions.

of the curves shows that if a patient belonging to typical progressors was initially on the lower green branch at $p_2 = 0.14$, then it is sufficient to reduce the value of p_2 to a value slightly less than 0.138, which will cause a spontaneous transition to the state depicted by the black line, characterized by a higher T cell concentration and lower viral load. It is then possible to increase the value of the parameter p_2 to the original value while staying on the same black line.

Hysteresis also occurs for parameters corresponding to the LTNP infection course, as demonstrated in Figure 3. The state depicted by the blue line at $p_2 = 0.155$ is stable, but it loses stability at p_2 smaller than 0.146. With further reduction of the parameter value, the system will move from a less favorable state (green branch) to a stable state with a higher concentration of $CD4^+$ T cells and lower viral load, depicted by the black solid line. After that, it is possible to return to the initial value of the parameter while remaining on this stable steady state branch.

Of practical importance is the question of the kinetics of the transition between different steady states when utilizing the hysteresis property. For a TP disease course, Figure 4 shows the transition dynamics from a less favorable state to a more favorable state for a system with hysteresis. It takes about 5,000 days to realize this transition with constant values of other system parameters. These results justify the relevance of further detailed study of such transitions.

Changes in steady states with a single administration of drugs

It is of independent interest to understand how the steady states of a system change under optimal control (Hadjiandreou et al., 2009). To this end, we investigated the time dependence of equilibria under therapeutic interventions $u_1(t)$, $u_2(t)$, which enter the right-hand sides of the model equations in the terms for the processes describing the infection of target cells and virus replication. Figure 5 shows the appearance of two new steady states at $t > 0.0005$, i. e., a change in the structure of the phase space of the model.

Figures 6–8 illustrate the steady-state changes when RDV, 3TC, and ZDV drugs are administered, the effects of which

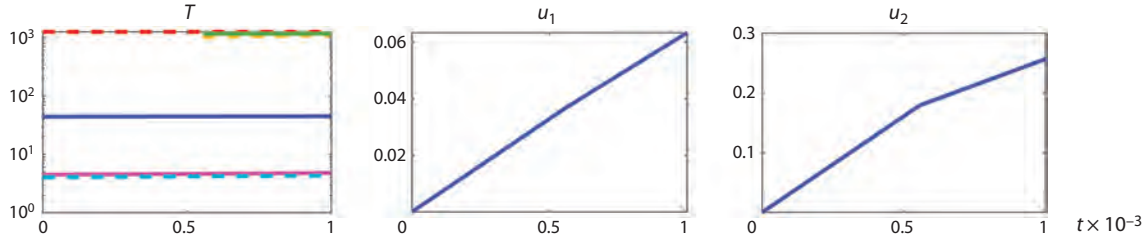


Fig. 5. Time dependencies t (in days) of the steady state variable T and control variables $u_1(t)$ and $u_2(t)$ at $0 \le t \le 0.001$ for long-term non-progression (LTNP).

Solid lines on the graph $T(t)$ correspond to stable steady states, dashed lines – to unstable ones.

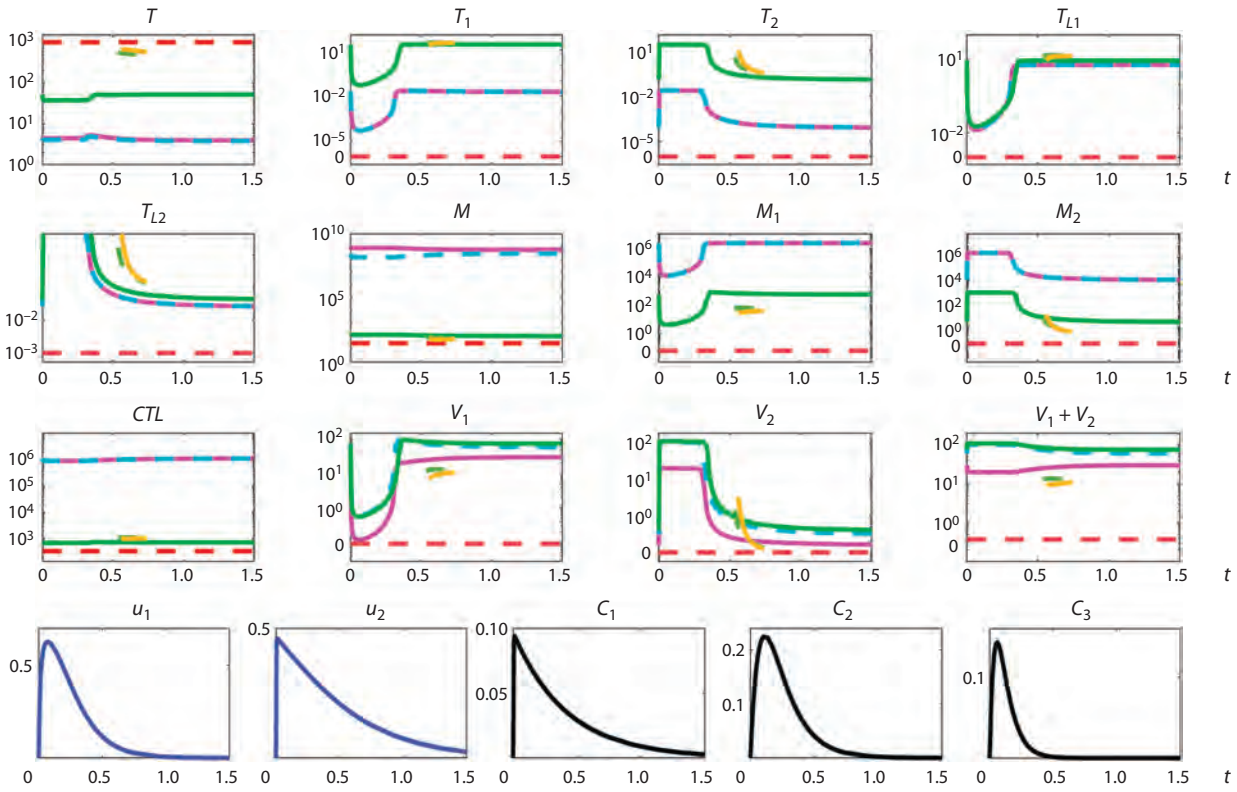


Fig. 6. Steady states and control actions for typical progression (TP) infection course.

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, different colors indicate different steady states. The horizontal axis shows time in days.

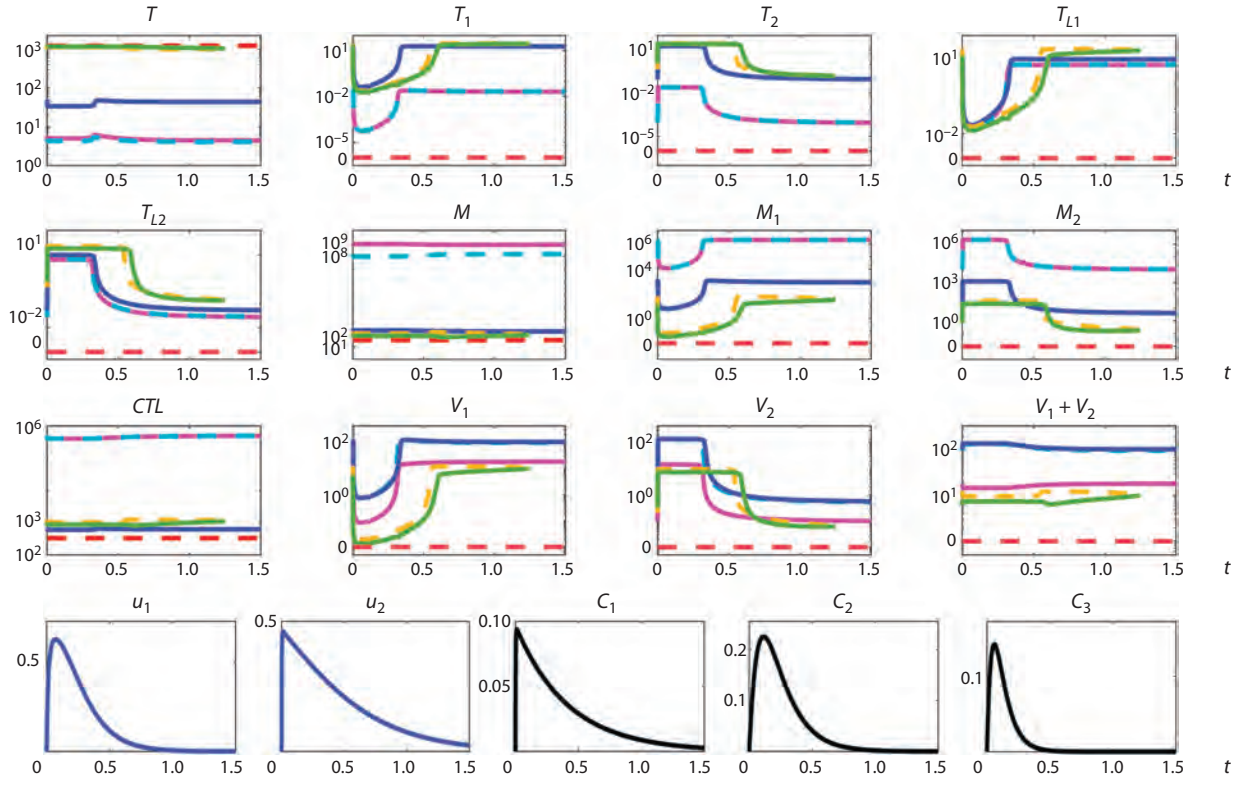


Fig. 7. Steady states of the model and control actions for long-term non-progressive flow (LTNP).

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, different colors indicate different steady states. The horizontal axis indicates time in days.

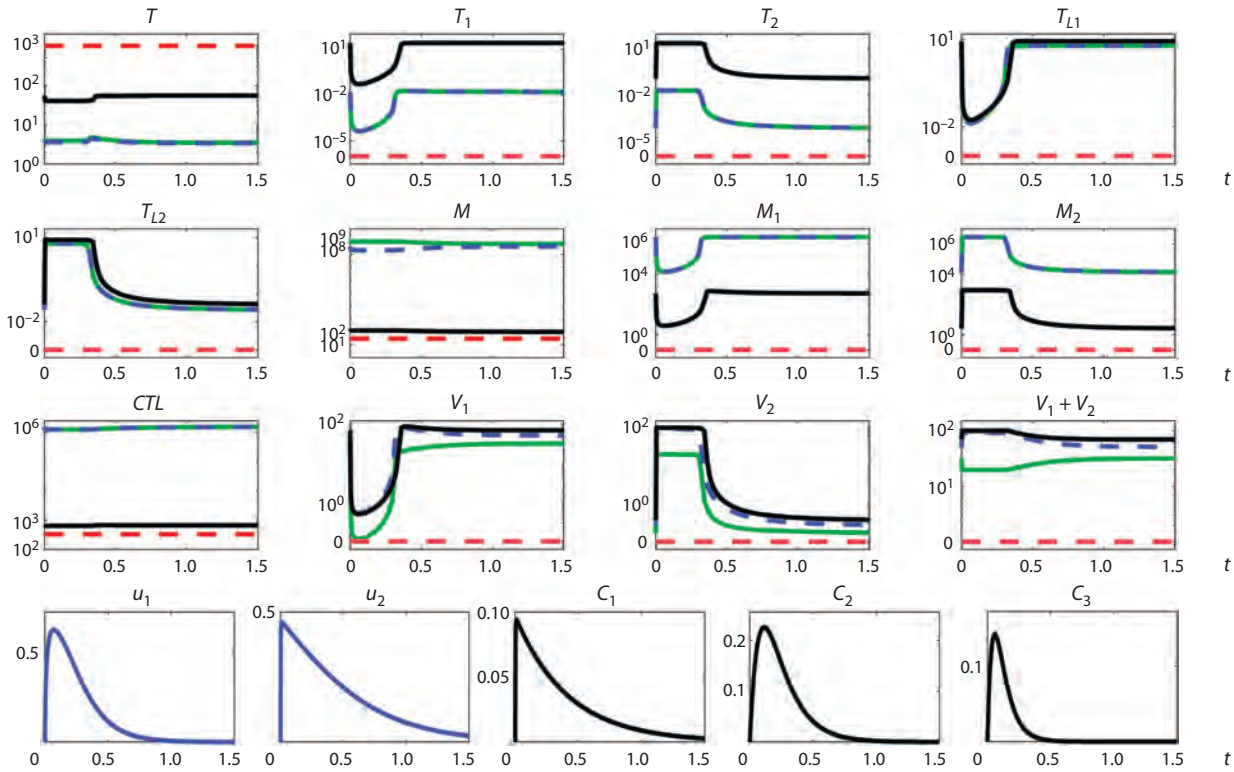


Fig. 8. Steady states of the model and control variables for rapid progression (RP) infection course.

Solid lines indicate stable steady states, dashed lines indicate unstable steady states, and different colors indicate different steady states. The horizontal axis indicates time in days.

are modeled using functions $C_1(t)$, $C_2(t)$, $C_3(t)$ through the control variables u_1 and u_2. The drugs are administered once at time $t = 0$. The solid lines indicate stable steady states and the dashed lines indicate unstable states, different colors indicate different steady states. The numerical results indicate that as the values of the control variables change, both stable and unstable steady states appear and then disappear. Thus, the application of optimal control methods leads to a change in the structure of the phase space of the model.

For all three variants of the course of HIV-1 infection, for one branch of the steady-state solutions, there is a short-term decrease in the values of variables characterizing the number of $CD4^+$ T cells and an increase in viral load due to an increase in the number of mutants and a decrease in steady-state concentrations of wild-type viruses. On the second stable branch, an opposite process takes place. In this case, in the case of a long-term non-progression course of HIV-1, the third branch of the stable equilibrium appears, which is characterized by a low viral load and, therefore, corresponds to more favorable dynamics. Thus, the impact of optimal control on the characteristics of equilibrium states depends essentially on the disease course phenotype (model parameters) and the neighborhood of the equilibrium in which the patient is in the case of bistability.

Thus, the response to the perturbation of the right-hand sides of the equations is qualitatively the same. The structure of the phase space changes, and as the control function impact is weakened, both stable and unstable steady states emerge and then disappear.

Discussion

A stable coexistence of the HIV-1 population and immune processes in the human body in various quantitative ratios is fundamentally important for the development of new strategies of HIV-1 therapy that belong to the category of functional treatment (cure) (Bocharov et al., 2022). In essence, it is the possibility of transferring the “virus–human host” system from a clinically more severe state to a milder infection stable steady state due to activation of immune defense mechanisms without further use of antiretroviral drugs that block viral replication. The presence of bi- or multistability indicates that by perturbing a certain trajectory of the system in the phase space, the transfer of the infectious disease to a more favorable regime can be accomplished. Both classical optimal control methods (Hadjiandreou et al., 2009; Bocharov et al., 2015) and our previously proposed methods based on optimal disturbances (Nechepurenko, Khristichenko, 2019; Khristichenko, Nechepurenko, 2022) exist as tools for constructing an appropriate control. Furthermore, there could be a case when a change in the kinetic parameters of biological and physiological processes is required to move the system into the region of bi- or multistability. The presence of hysteresis allows one to develop treatment approaches that utilize temporary parametric shifts with subsequent return to the initial values of the changed parameters. The identified properties of the mathematical model of HIV-1 infection, which has a fairly typical structure, theoretically confirm the potential feasibility of corresponding combination immune-based therapeutic interventions (Landovitz et al., 2023).

The obtained estimates of the parameter regions enabling the existence of bistability together with the characteristics of bifurcation diagrams show that as the severity of the HIV-1 infection increases, i. e. in the transition from long-term non-progressor to typical progressor and further to rapid progressor phenotype, the range of values of the activation rate of innate immunity cells, at which the bistability takes place, increases. Meanwhile, the properties of bifurcation diagrams also change. These specific features of the response of an HIV-infected patient should be taken into account and used in the design of immunomodulatory regiments.

Finally, we showed that the impact of optimal control on the characteristics of equilibria depends significantly on the phenotype of HIV-1 infection (determined by system parameters) and the neighborhood of the equilibrium in which the patient is located in the case of bi- or multistability.

Conclusion

In this paper, we have computed and numerically analyzed the steady states of the mathematical model of HIV-1 infection for sets of parameters corresponding to phenotypically different variants of the course of the infection: typical progression, long-term non-progression and rapid progression. The results of the bifurcation analysis of the HIV-1 infection model indicate that implementation of an effective functional cure of infected patients requires the development of a personalized approach that takes into account both the properties of the HIV-1 quasispecies population and the patient's immune status. Overall, our study forms a theoretical basis for the development of combination immune-based therapy of HIV-1 infected patients.

References

Akın E., Yeni G., Perelson A.S. Continuous and discrete modeling of HIV-1 decline on therapy. *J. Math. Biol.* 2020;81(1):1-24. DOI 10.1007/s00285-020-01492-z

Banks H.T., Hu S., Rosenberg E. A dynamical modeling approach for analysis of longitudinal clinical trials in the presence of missing endpoints. *Appl. Math. Lett.* 2017;63:109-117. DOI 10.1016/j.aml.2016.07.002

Bocharov G., Chereshnev V., Gainova I., Bazhan S., Bachmetyev B., Argilaguet J., Martinez J., Meyerhans A. Human immunodeficiency virus infection: from biological observations to mechanistic mathematical modelling. *Math. Model. Nat. Phenom.* 2012;7(5):78-104. DOI 10.1051/mmnp/20127507

Bocharov G., Kim A., Krasovskii A., Chereshnev V., Glushenkova V., Ivanov A. An extremal shift method for control of HIV infection dynamics. *Russ. J. Numer. Anal. Math. Model.* 2015;30(1):11-25. DOI 10.1515/rnam-2015-0002

Bocharov G.A., Nechepurenko Y.M., Khristichenko M.Y., Grebennikov D.S. Optimal perturbations of systems with delayed independent variables for control of dynamics of infectious diseases based on multicomponent actions. *J. Math. Sci*. 2021;253(5):618-641. DOI 10.1007/s10958-021-05258-w

Bocharov G., Grebennikov D., Cebollada Rica P., Domenjo-Vila E., Casella V., Meyerhans A. Functional cure of a chronic virus infection by shifting the virus – host equilibrium state. *Front. Immunol*. 2022;13:904342. DOI 10.3389/fimmu.2022.904342

Gandhi R.T., Bedimo R., Hoy J.F., Landovitz R.J., Smith D.M., Eaton E.F., Lehmann C., Springer S.A., Sax P.E., Thompson M.A., Benson C.A., Buchbinder S.P., Del Rio C., Eron J.J., Jr., Günthard H.F., Molina J.-M., Jacobsen D.M., Saag M.S. Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2022 recommendations of the International Antiviral Society-USA Panel. *JAMA*. 2023;329(1):63-84. DOI 10.1001/jama.2022.22246

Geddes K.O., Czapor S.R., Labahn G. Algorithms for Computer Algebra. Boston: Kluwer Academic, 1992

Golub G.H., Van Loan C.F. Matrix Computations. Baltimore: Johns Hopkins Univ. Press, 1989

Grossman Z., Singh N.J., Simonetti F.R., Lederman M.M., Douek D.C., Deeks S.G., Kawabe T., Bocharov G., Meier-Schellersheim M., Alon H., Chomont N., Grossman Z., Sousa A.E., Margolis L., Maldarelli F. "Rinse and replace": boosting T cell turnover to reduce HIV-1 reservoirs. *Trends Immunol.* 2020;41(6):466-480. DOI 10.1016/j.it.2020.04.003

Hadjiandreou M.M., Conejeros R., Wilson I. HIV treatment planning on a case-by-case basis. *Int. J. Bioeng. Life Sci.* 2009;3(8):387-396

Hairer E., Nørsett S.P., Wanner G. Solving Ordinary Differential Equations I. Springer Series in Computational Mathematics. Vol. 8. Berlin: Springer, 1987. DOI 10.1007/978-3-662-12607-3

Joly M., Pinto J.M. Role of mathematical modeling on the optimal control of HIV-1 pathogenesis. *AIChE J.* 2006;52(3):856-884. DOI 10.1002/aic.10716

Khristichenko M.Y., Nechepurenko Y.M. Computation of periodic solutions to models of infectious disease dynamics and immune response. *Russ. J. Numer. Anal. Math. Model.* 2021;36(2):87-99. DOI 10.1515/rnam-2021-0008

Khristichenko M.Y., Nechepurenko Y.M. Optimal disturbances for periodic solutions of time-delay differential equations. *Russ. J. Numer. Anal. Math. Model.* 2022;37(4):203-212. DOI 10.1515/rnam-2022-0017

Khristichenko M.Yu., Nechepurenko Yu.M., Grebennikov D.S., Bocharov G.A. Numerical analysis of stationary solutions of systems with delayed argument in mathematical immunology. *Sovremennaya Matematika. Fundamental'nye Napravleniya = Contemporary Mathematics. Fundamental Directions*. 2022;68(4):686-703. DOI 10.22363/2413-3639-2022-68-4-686-703 (in Russian)

Khristichenko M., Nechepurenko Y., Grebennikov D., Bocharov G. Numerical study of chronic hepatitis B infection using Marchuk–Petrov model. *J. Bioinform. Comput. Biol*. 2023;21(2):2340001. DOI 10.1142/S0219720023400012

Landovitz R.J., Scott H., Deeks S.G. Prevention, treatment and cure of HIV infection. *Nat. Rev. Microbiol*. 2023;21(10):657-670. DOI 10.1038/s41579-023-00914-1

Ludewig B., Stein J.V., Sharpe J., Cervantes-Barragan L., Thiel V., Bocharov G. A global "imaging" view on systems approaches in immunology. *Eur. J. Immunol.* 2012;42(12):3116-3125. DOI 10.1002/eji.201242508

Nechepurenko Y.M., Khristichenko M.Y. Computation of optimal disturbances for delay systems. *Comput. Math. and Math. Phys.* 2019;59(5):731-746. DOI 10.1134/S0965542519050129

Nechepurenko Y., Khristichenko M., Grebennikov D., Bocharov G. Bistability analysis of virus infection models with time delays. *Discrete Cont. Dyn. Syst. - S*. 2020;13(9):2385-2401. DOI 10.3934/dcdss.2020166

Niessl J., Baxter A.E., Mendoza P., Jankovic M., Cohen Y.Z., Butler A.L., Lu C.-L., Dubé M., Shimeliovich I., Gruell H., Klein F., Caskey M., Nussenzweig M.C., Kaufmann D.E. Combination anti-HIV-1 antibody therapy is associated with increased virus-specific T cell immunity. *Nat. Med*. 2020;26(2):222-227. DOI 10.1038/s41591-019-0747-1

Nowak M.A., May R.M. Virus Dynamics: Mathematical Principles of Immunology and Virology. Oxford: Oxford Univ. Press, 2000

Perelson A.S., Nelson P.W. Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev.* 1999;41(1):3-44. DOI 10.1137/S0036144598335107

Rasmussen T.A., Søgaard O.S. Clinical interventions in HIV cure research. In: Zhang L., Lewin S.R. (Eds.) HIV Vaccines and Cure. Advances in Experimental Medicine and Biology. Vol. 1075. Singapore: Springer, 2018;285-318. DOI 10.1007/978-981-13-0484-2_12

Savinkova A.A., Savinkov R.S., Bakhmetyev B.A., Bocharov G.A. Mathematical modeling and control of HIV infection dynamics taking into account hormonal regulation. *Vestnik Rossiyskogo Universiteta Druzhby Narodov. Seriya Meditsina = RUDN Journal of Medicine*. 2019;23(1):79-103. DOI 10.22363/2313-0245-2019-23-1-79-103 (in Russian)

Trickey A., Zhang L., Gill M.J., Bonnet F., Burkholder G., Castagna A., Cavassini M., Cichon P., Crane H., Domingo P., Grabar S., Guest J., Obel N., Psichogiou M., Rava M., Reiss P., Rentsch C.T., Riera M., Schuettfort G., Silverberg M.J., Smith C., Stecher M., Sterling T.R., Ingle S.M., Sabin C.A., Sterne J.A.C. Associations of modern initial antiretroviral drug regimens with all-cause mortality in adults with HIV in Europe and North America: a cohort study. *Lancet HIV*. 2022;9(6):e404-e413. DOI 10.1016/S2352-3018(22)00046-7

Villani A.-C., Sarkizova S., Hacohen N. Systems immunology: learning the rules of the immune system. *Annu. Rev. Immunol.* 2018;36(1):813-842. DOI 10.1146/annurev-immunol-042617-053035

ORCID ID

G.A. Bocharov orcid.org/0000-0002-5049-0656

Acknowledgements. This work was financially supported by the Russian Science Foundation, project No. 22-71-10028.

Conflict of interest. The authors declare no conflict of interest.

Received July 14, 2023. Revised September 15, 2023. Accepted September 19, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):768-775
DOI 10.18699/VJGB-23-89

Original Russian text https://vavilovj-icg.ru/

Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium

V.A. Ivanisenko[1, 2, 7 *], N.V. Basov[2, 3 *], A.A. Makarova[1], A.S. Venzel[1, 7], A.D. Rogachev[2, 3] ✉, P.S. Demenkov[1, 2, 7], T.V. Ivanisenko[1, 2, 7], M.A. Kleshchev[1], E.V. Gaisler[2, 3], G.B. Moroz[4], V.V. Plesko[4], Y.S. Sotnikova[2, 3, 5], Y.V. Patrushev[2, 5], V.V. Lomivorotov[4, 6], N.A. Kolchanov[1, 2], A.G. Pokrovsky[2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] N.N. Vorozhtsov Novosibirsk Institute of Organic Chemistry of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] E. Meshalkin National Medical Research Center of the Ministry of Health of Russian Federation, Novosibirsk, Russia
[5] Boreskov Institute of Catalysis of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[6] Penn State Health Milton S. Hershey Medical Center, Hershey, PA, USA
[7] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
✉ rogachev@nioch.nsc.ru

Abstract. Postoperative delirium (POD) is considered one of the most severe complications, resulting in impaired cognitive function, extended hospitalization, and higher treatment costs. The challenge of early POD diagnosis becomes particularly significant in cardiac surgery cases, as the incidence of this complication exceeds 50 % in certain patient categories. While it is known that neuroinflammation, neurotransmitter imbalances, disruptions in neuroendocrine regulation, and interneuronal connections contribute significantly to the development of POD, the molecular, genetic mechanisms of POD in cardiac surgery patients, along with potential metabolomic diagnostic markers, remain inadequately understood. In this study, blood plasma was collected from a group of patients over 65 years old after cardiac surgery involving artificial circulation. The collected samples were analyzed for sphingomyelin content and quantity using high-performance liquid chromatography coupled with mass spectrometry (HPLC-MS/MS) methods. The analysis revealed four significantly different sphingomyelin contents in patients with POD compared to those who did not develop POD (control group). Employing gene network reconstruction, we perceived a set of 82 regulatory enzymes affiliated with the genetic coordination of the sphingolipid metabolism pathway. Within this set, 47 are assumed to be regulators of gene expression, governing the transcription of enzymes pivotal to the metabolic cascade. Complementing this, an additional assembly of 35 regulators are considered to be regulators of activity, degradation, and translocation dynamics of enzymes integral to the aforementioned pathway. Analysis of the overrepresentation of diseases with which these regulatory proteins are associated showed that the regulators can be categorized into two groups, associated with cardiovascular pathologies (CVP) and neuropsychiatric diseases (NPD), respectively. The regulators associated with CVP are expectedly related to the effects on myocardial tissue during surgery. It is hypothesized that dysfunction of NPD-associated regulators may specifically account for the development of POD after cardiac surgery. Thus, the identified regulatory genes may provide a basis for planning further experiments, in order to study disorders at the level of expression of these genes, as well as impaired function of proteins encoded by them in patients with POD. The identified significant sphingolipids can be considered as potential markers of POD.
Key words: LC-MS/MS; metabolomics; lipidomics; postoperative delirium; cardiac surgery; biomarkers; sphingolipids; gene networks; ANDSystem.

For citation: Ivanisenko V.A., Basov N.V., Makarova A.A., Venzel A.S., Rogachev A.D., Demenkov P.S., Ivanisenko T.V., Kleshchev M.A., Gaisler E.V., Moroz G.B., Plesko V.V., Sotnikova Y.S., Patrushev Y.V., Lomivorotov V.V., Kolchanov N.A., Pokrovsky A.G. Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):768-775. DOI 10.18699/VJGB-23-89

Применение генных сетей к анализу результатов метаболомного скрининга плазмы крови пациентов с послеоперационным делирием

В.А. Иванисенко[1, 2, 7 *], Н.В. Басов[2, 3 *], А.А. Макарова[1], А.С. Вензель[1, 7], А.Д. Рогачев[2, 3] ✉, П.С. Деменков[1, 2, 7], Т.В. Иванисенко[1, 2, 7], М.А. Клещев[1], Е.В. Гайслер[2, 3], Г.Б. Мороз[4], В.В. Плеско[4], Ю.С. Сотникова[2, 3, 5], Ю.В. Патрушев[2, 5], В.В. Ломиворотов[4, 6], Н.А. Колчанов[1, 2], А.Г. Покровский[2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

* Authors contributed equally to the study.

[3] Новосибирский институт органической химии им. Н.Н. Ворожцова Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Национальный медицинский исследовательский центр им. академика Е.Н. Мешалкина Министерства здравоохранения Российской Федерации, Новосибирск, Россия
[5] Федеральный исследовательский центр «Институт катализа им. Г.К. Борескова Сибирского отделения Российской академии наук», Новосибирск, Россия
[6] Медицинский центр им. Милтона Херши, Херши, Пенсильвания, США
[7] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
rogachev@nioch.nsc.ru

Аннотация. Послеоперационный делирий (ПОД) является серьезным осложнением, приводящим к нарушению когнитивных функций пациентов, увеличению длительности госпитализации, а также повышению расходов на лечение пациента. Проблема ранней диагностики ПОД приобретает особую важность в случае кардиохирургических операций, поскольку частота развития такого осложнения у некоторых категорий пациентов превышает 50 %. Известно, что в развитие ПОД большой вклад вносят нейровоспаление, дисбаланс нейромедиаторов, нарушение нейроэндокринной регуляции и межнейрональных связей, однако молекулярно-генетические механизмы ПОД у пациентов, перенесших кардиохирургические операции, а также метаболомные диагностические маркеры, до сих пор плохо изучены. В данной работе с помощью метода высокоэффективной жидкостной хроматографии с масс-спектрометрической детекцией (ВЭЖХ-МС/МС) был проведен анализ содержания ряда сфингомиелинов в плазме крови пациентов старше 65 лет, взятой после операции на сердце в условиях искусственного кровообращения. Найдено четыре статистически значимо различающихся по содержанию сфингомиелина у пациентов с ПОД по сравнению с пациентами, у которых не развился ПОД (контрольная группа). С помощью реконструкции генных сетей, описывающих генетическую регуляцию пути метаболизма сфинголипидов, определены 82 регуляторных белка, из которых 47 – регуляторы экспрессии генов, кодирующих ферменты метаболического пути, и 35 – регуляторы активности, деградации и транспорта ферментов данного пути. Анализ перепредставленности заболеваний, с которыми ассоциированы эти регуляторные белки, показал, что регуляторы можно разбить на две группы, ассоциированные с сердечно-сосудистыми патологиями и с нервно-психическими заболеваниями соответственно. Регуляторы, ассоциированные с сердечно-сосудистыми патологиями, ожидаемо связаны с воздействием на ткани миокарда во время операции. Сделано предположение, что нарушение функции регуляторов, ассоциированных с нервно-психическими заболеваниями, может специфически обусловливать развитие ПОД после кардиохирургической операции. Таким образом, выявленные регуляторные гены могут представлять основу для планирования дальнейших экспериментов по изучению нарушений на уровне экспрессии данных генов, а также нарушения функции кодируемых ими белков у пациентов с ПОД. Идентифицированные значимые сфинголипиды могут рассматриваться как потенциальные маркеры послеоперационного делирия.
Ключевые слова: ВЭЖХ-МС/МС; метаболомика; липидомика; послеоперационный делирий; кардиохирургия; биомаркеры; сфинголипиды; генные сети; ANDSystem.

Introduction

Postoperative delirium (POD) is a serious complication of the early postoperative period. Its incidence in cardiovascular surgery is 52 % (Brown, 2014). The development of POD leads to a worse prognosis, including longer hospitalization duration, increased complications and mortality, impaired cognitive function and physical status, and increased patient costs (Pisani et al., 2009; Gottesman et al., 2010). Delirium and postoperative cognitive impairment most commonly occur in patients over 60 years of age (Morimoto et al., 2009). Factors such as CNS hypoxia, embolism, neurotransmitter release, systemic inflammatory responses and other disorders, including metabolic issues, contribute to this phenomenon (Wimmer-Greinecker et al., 1998; Cerejeira et al., 2010).

Metabolomics is a branch of bioanalytical chemistry focused on the identification and quantification of low molecular weight metabolites (<1,500 Da). The metabolomic approach can be used to search for associations between metabolic signatures and disease phenotypes. In particular, metabolomic methods allow the detection of low molecular weight metabolites capable of crossing the blood-brain barrier, making metabolomic analysis a powerful tool for identifying markers of delirium (Ke et al., 2019). For example, several studies have shown that disturbances in energy metabolism, amino acid biosynthesis, omega-3 and omega-6 fatty acid deficiency, and glutamate-glutamine cycle dysfunction are associated with postoperative delirium in non-cardiac surgery (Guo et al., 2019; Tripp et al., 2021).

Previously, a number of our studies have shown the possibility of using the results of metabolomic screening in the search for biomarkers of pathologies, as well as the reconstruction of gene networks based on the obtained data. Thus, using statistical analysis of metabolomic profiles of cerebrospinal fluid (CSF) and blood plasma of patients with high-grade glioma obtained by HPLC-MS/MS, we have revealed correlations between metabolomic profiles of blood plasma and CSF (Rogachev et al., 2021). Metabolomic analysis combined with gene network reconstruction using ANDSystem to interpret metabolomic data (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022) allowed us to identify key SARS-CoV-2 proteins whose interactions with human proteins could lead to impaired metabolic processes in COVID-19 patients (Ivanisenko V.A. et al., 2022).

Sphingomyelins (SM) are among the major phospholipids that make up the hydrophobic matrix of plasma membranes of mammalian cells; however, in response to stress, sphingomyelins can be cleaved by sphingomyelinase into phosphatidylcholine and ceramide, which have a signaling function. Changes in sphingomyelin metabolism can affect the balance of neurotransmitters in the brain, cause disruption of

neuronal connectivity, and induction of neuroinflammation, making them an important target for studying the mechanisms of delirium pathogenesis (Wang, Shen, 2018; Xiao et al., 2023).

In this study, the content of 9 phospholipids belonging to the sphingomyelin class in the plasma of patients undergoing cardiac surgery was analyzed using HPLC-MS/MS. There were 4 statistically significant different sphingomyelin contents in patients with POD compared with patients who did not develop POD (control group).

To explain possible mechanisms of sphingolipid metabolism disorders, we reconstructed gene networks describing the genetic regulation of the KEGG pathway "Sphingolipid metabolism" (hsa: 00600) using ANDSystem. Analysis of gene networks allowed us to identify 35 regulators of transport, activity and degradation of enzymes of this pathway, as well as 47 regulators of expression of genes encoding these enzymes.

Materials and methods

Patients. The study included patients over 65 years of age who underwent cardiac surgery under artificial circulation. Exclusion criteria were: emergency intervention, aortic surgery, hemodynamically significant carotid artery stenosis, Parkinson's disease, liver cirrhosis (Child-Pugh B or C), taking anticholinergic drugs, antidepressants, antiepileptic and chemotherapeutic drugs. Patients were recruited from June 2019 to January 2021. A total of 39 patients were included in the study (Table 1). Within 5 days after surgery, patients were evaluated for postoperative delirium using the CAM-ICU (Confusion Assessment Method for the Intensive Care Unit) test. The first test was performed 6–8 hours after surgery, and then the patients were assessed twice a day. Delirium was considered to be present if the CAM-ICU test was positive at least once.

The study was approved by the Ethics Committee of the E. Meshalkin National Medical Research Center (Novosibirsk, Russia).

Blood sampling and sample preparation. Blood samples were collected from patients 24 hours after cardiac surgery. Venous blood was collected into 10 mL BD Vacutainer® KEDTA tubes containing potassium EDTA as anticoagulant. Plasma was separated from blood cells by centrifugation for 15 min at 2,000 g and +4 °C, separated into aliquots and stored frozen at –80 °C until further use.

All samples were processed simultaneously according to the protocol described in (Li et al., 2017): 400 μL of a chilled mixture of methanol and acetonitrile (1:1) was added to 100 μL of blood plasma. The samples were shaken on a shaker, then centrifuged for 15 min at +4 °C and 16,000 rpm. The supernatant was transferred to a glass vial insert and analyzed. Two quality control samples, obtained by mixing equal volumes of plasma samples from patients with POD and controls, were prepared using the same procedure.

HPLC-MS/MS analysis was performed on a Shimadzu LC-20AD Prominence chromatograph equipped with a gradient pump, SIL-20AC autosampler (Shimadzu, Japan), thermostated at 10 °C, and CTO-10ASvp column thermostat, with a temperature of 35 °C. Chromatographic separation was carried out on a monolithic column with 1-vinyl-1,2,4-triazole based sorbent (Basov et al., 2024). The monolithic material was synthesized in glass tubes with an inner diameter of 2 mm as described previously (Patrushev et al., 2020). Mobile phase A was an aqueous 20 mM ammonium carbonate solution adjusted to pH = 9.8 with 25 % aqueous ammonia solution and containing 5 vol % acetonitrile; mobile phase B was pure acetonitrile. The elution gradient was as follows: 0 min – 0 % B, 1 min – 0 % B, 6 min – 98 % B, 16 min – 98 % B, after which the column was equilibrated for 3 min. The flow rate was 300 μL/min and the sample volume was 2 μL.

Table 1. Sex and age characteristics of patients

Group	Sex (M/F)	Age, yrs				
		Min	Max	Mean	Median	Standard deviation
Control	11/16	65	75	69.6	70	3.0
POD	5/7	65	79	69.7	69.5	4.3

Metabolites were detected on an API 6500 QTRAP mass spectrometer (AB SCIEX, USA) equipped with an electrospray ionization source operating in positive ionization mode. Metabolites were detected in multiple reaction monitoring (MRM) mode.

The main mass spectrometric parameters were as follows. The voltage at the ion source was 5500 V. The dryer gas temperature was 475 °C, CAD gas pressure was "high", GS1, GS2 and curtain gas pressures were 33, 33 and 30 psi, respectively. The declustering potential (DP) was 91 V, the entry potential (EP) was 10 V, and the collision cell exit potential (CXP) was 10 V. The precursor and fragment ion transitions, metabolite names, residence times, and corresponding collision energies are presented in the Supplementary Table S1[1]. Instrument control and data acquisition were performed using Analyst 1.6.3 software (AB SCIEX, Framingham, MA, USA). Chromatograms were processed using the MultiQuant 2.1 program (AB SCIEX, USA).

Data preprocessing and statistics analysis. The raw data were preprocessed to fill in missing values of metabolite content in the analyzed samples as follows. If the number of samples with missing values did not exceed 5 % of the total number of values for 39 patients, the median was calculated for the remaining samples and taken as the metabolite content value. This approach is due to the robustness of the median to outliers. Statistical differences in the content of metabolites in blood plasma samples in the group of patients with and without POD were assessed using the nonparametric Mann–Whitney criterion.

Reconstruction and analysis of gene networks. The list of genes encoding enzymes involved in the "Sphingolipid metabolism" pathway (ID: hsa00600) was extracted from the KEGG database (https://www.kegg.jp/kegg/pathway.html, Kanehisa, 2002; Kanehisa et al., 2022). The regulatory gene network was reconstructed using the ANDSystem software and information system (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022). Work with the ANDSystem knowledge base was performed using the ANDVisio program

[1] Supplementary Tables S1–S10 are available at: https://vavilovj-icg.ru/download/pict-2023-27/appx24.xlsx

module. Analysis of overrepresentation of biological processes (Gene Ontology) and diseases associated with regulatory gene network proteins was performed using the web-based tool DAVID (https://david.ncifcrf.gov/tools.jsp, Huang D.W. et al., 2009).

Results

Study of sphingolipid content in patients' blood plasma using the HPLC-MS/MS method

Since impaired sphingomyelin metabolism may contribute to the development of delirium, the aim of our analysis was to investigate their role in the complication of POD by examining their content in the plasma of patients after cardiac surgery. Specifically, we comparatively analyzed SM expression in the plasma of patients undergoing cardiac surgery. The metabolites of this class that had a significant statistical difference in content within the samples taken from the group of patients with POD were compared with those from the group of patients who did not develop POD and are summarized in Table 2.

According to the Mann–Whitney test, out of these 9 sphingomyelins analyzed, four (SM(d18:1/22:2 OH), SM(d18:1/24:0), SM(d18:1/24:1), and SM(d18:1/22:2)) showed statistically significant (p-value < 0.05) differences between the studied patient groups. We hypothesized that the impaired metabolism of sphingolipids may be related to the impaired metabolic pathway of their biosynthesis. To test this hypothesis, using the ANDSystem software and information system, we reconstructed and analyzed the gene network describing the regulation of expression of genes encoding enzymes of the KEGG "Sphingolipid metabolism" metabolic pathway, as well as the regulation of transport, activity, and degradation of these enzymes.

Reconstruction of the regulatory gene network

To reconstruct the regulatory gene network, a list of genes encoding enzymes involved in sphingolipid metabolism "Sphingolipid metabolism" (hsa00600) was extracted from the KEGG database. The resulting list contained 43 human genes (Supplementary Table S2). The gene network graph was reconstructed in the "Query Wizard" module of ANDVisio.

It should be noted that in the gene network we considered only regulatory connections directed from regulatory proteins to enzymes of the metabolic pathway. The resulting gene network contained 43 human genes, 125 proteins (43 metabolic pathway enzymes and 82 regulatory proteins), and 159 interactions between them (see the Figure). Different types of interactions between gene network members were represented in the following ratio: 28 links corresponding to the type "regulation of activity", 2 – "regulation of degradation", 4 – "proteolysis", 8 – "regulation of transport", 43 – "expression", 74 – "regulation of expression".

To investigate the association of regulatory proteins with pathologies, we analyzed the overrepresentation of diseases and biological processes by Gene Ontology, using the web-based tool DAVID. A list consisting of 82 genes encoding gene-network regulatory proteins was provided as the input. The results of the disease and biological processes overrepresentation analysis are summarized in Supplementary Tables S3 and S4, respectively.

Table 2. Statistical significance of differences between the group of patients with POD and the control group in the content of metabolites in blood plasma samples when compared by the Mann–Whitney test

Metabolites	p-value
SM(d18:1/22:2 OH)	0.0273
SM(d18:1/24:0)	0.0430
SM(d18:1/24:1)	0.0462
SM(d18:1/22:2)	0.0496
SM(d18:1/18:0)	0.0750
SM(d18:1/22:1)	0.1483
SM(d18:1/20:1)	0.3693
SM(d18:1/20:0)	0.5129
SM(d18:1/24:2)	0.5943

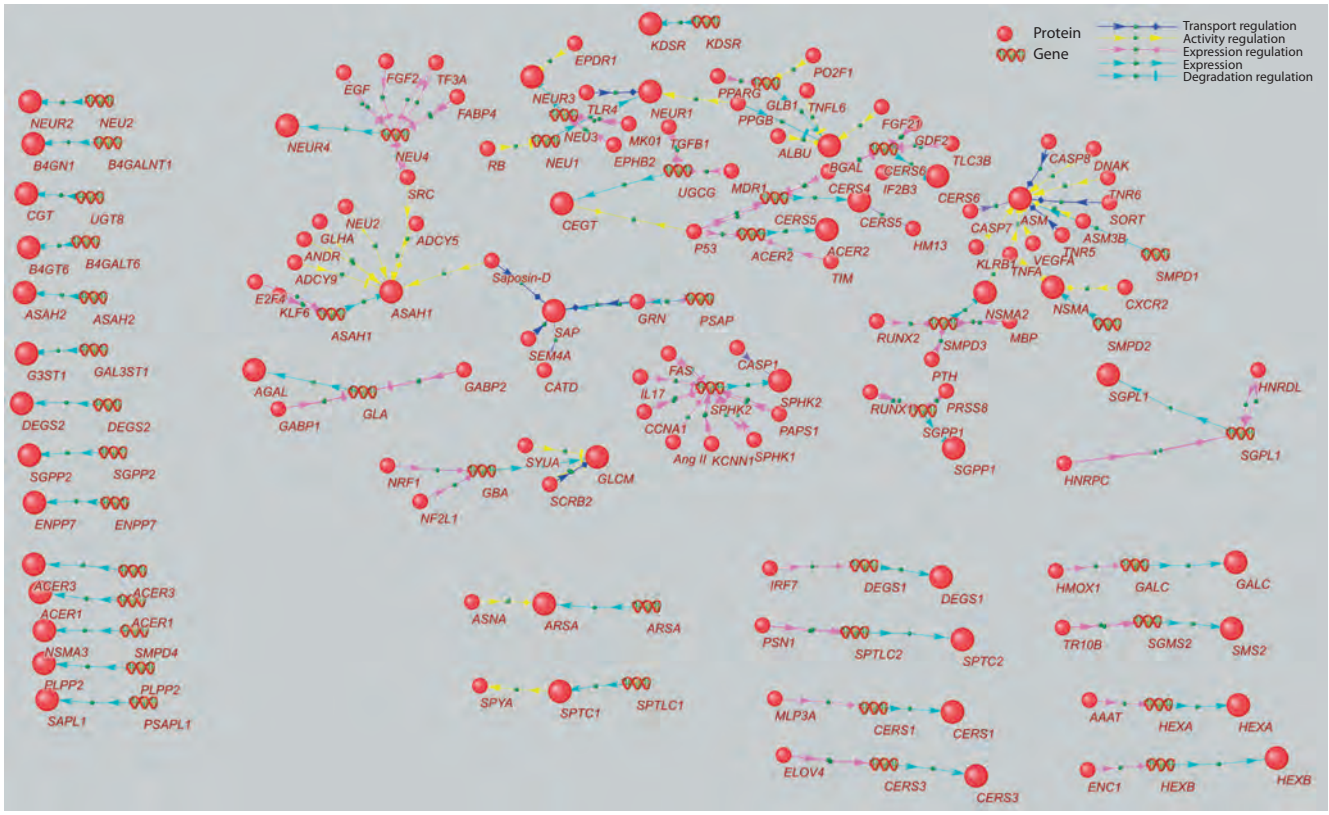
All regulatory proteins represented in the gene network (see the Figure) can be divided into two groups: (1) regulators of gene expression and (2) regulators of activity, stability, transport, etc., which can be called regulators of protein function. To investigate the features associated with these diseases and biological processes, overrepresentation analysis was performed separately for each of these groups of proteins (Supplementary Tables S5–S8).

Discussion

According to the literature, sphingomyelins (SM) play an important role in nervous system function, and alterations in their metabolism may contribute to the development of delirium by inducing neuroinflammation, altering neurotransmitter balance, and disrupting neuronal connectivity (Wang, Shen, 2018; Xiao et al., 2023). Our metabolomic analysis using HPLC-MS/MS of blood plasma from patients undergoing cardiac surgery allowed us to identify 4 out of 9 sphingomyelins, the content of which had a significant statistical difference in the analyzed samples of patients with POD compared to patients who did not develop POD (see Table 2).

To study potential mechanisms of sphingolipid metabolism disorders, a gene network (see the Figure) describing the regulation of gene expression and function of the enzymes encoded by them, participants of the KEGG metabolic pathway "Sphingolipid metabolism" (Sphingolipid metabolism, hsa: 00600), was reconstructed using ANDSystem. Network analysis showed that 82 regulatory proteins were involved in the regulation of the metabolic pathway, the dysfunction of which could influence the impairment of sphingolipid metabolism. Based on the enrichment analysis of the list of genes encoding these proteins with disease-associated genes, 168 statistically significantly overrepresented diseases were identified.

To simplify the presentation of the results, the list of diseases was divided into five groups (Table 3). The most significant disease was from the group of cardiovascular system pathologies, which may be due to the fact that all patients underwent cardiac surgery due to cardiac pathologies. The surgeries and medical procedures performed, such as



Gene network regulation of the sphingolipid metabolism pathway.

artificial circulation, may also explain the presence of the groups "inflammation", "renal pathologies" and "surgical intervention" among the identified significant pathologies (Stafford-Smith et al., 2008; Squiccimarro et al., 2019). These pathologies could be concurrently associated with both groups of patients, with and without POD, as each had undergone cardiac surgery.

The "neuropsychiatric diseases" group of pathologies is of particular interest in the context of the development of postoperative delirium. In particular, H. Huang et al. (2022) discuss the role of neuroinflammation in the development of postoperative delirium. The authors emphasize neuroinflammation and disruption of the blood-brain barrier as some of the main pathophysiological factors in the onset of delirium. The association of neuropsychiatric pathologies with POD is also widely discussed in the scientific literature. For example, O'Sullivan et al. consider that the link between delirium and depressive disorder may be due to common pathophysiological mechanisms including impaired stress and inflammatory responses, monoamine and melatoninergic signaling (O'Sullivan et al., 2014). According to our analysis, at the molecular genetic level, these pathophysiologic mechanisms may involve genetic regulation of the sphingolipid metabolism pathway. The list of regulatory genes from the gene network associated with the "neuropsychiatric diseases" group is summarized in Supplementary Table S9.

Statistical analysis of Gene Ontology overrepresentation of biological processes based on the list of regulatory genes identified 67 significant biological processes (BPs, Supplementary Table S4). The list of biological processes was categorized into 7 groups to represent the results (Table 4). Among the significant BPs were found fundamental regulatory processes, including regulation of transcription, regulation of proliferation, activation of cell signaling pathways, etc. These results were expected because the gene network members are regulators of gene expression and functions of the enzymes they encode. The set of regulators we analyzed turned out to be enriched with genes involved in the process of cardiovascular cell proliferation (see Table 4 and Supplementary Table S10), which can be explained by the activation of regenerative processes after surgical intervention. In addition to those mentioned above, it is possible to identify more specifically delirium-related BPs, such as inflammatory processes, response to stress factors and regulation of apoptosis (Steiner, 2011; Vutskits, Xie, 2016).

Notably, the regulatory connections in the gene network can be divided into two groups: regulators of the gene expression and regulators of the functions (activity, degradation, and transport) of protein products of their expression. It is of interest to consider whether there are characteristic features related to the molecular mechanisms of delirium development for regulators from these two separate groups. We analyzed the overrepresentation of diseases and biological processes separately for regulators of expression as well as for regulators of protein function (see Supplementary Tables S5–S8).

Unexpectedly, when considering regulators of protein function, the top ten most significant pathologies were neuropsychiatric diseases (e. g., schizophrenia, bipolar disorders, autism), which, according to the literature, are

Table 3. Statistical significance of disease overrepresentation based on gene-regulatory list analysis

Pathology group	Number of pathologies in the group	The most significant pathology	FDR	Number of genes
Pathologies of the cardiovascular system	23	Hypertension	7.7×10^{-6}	12
Kidney pathologies	8	Acute renal failure	7.3×10^{-6}	10
Inflammatory processes	4	Inflammation	7.3×10^{-6}	11
Operative intervention	2	Reperfusion injury	1.4×10^{-5}	9
Neuropsychiatric diseases	26	Depressive disorder	2.1×10^{-4}	12

Note. False Discovery Rate (FDR) and the number of genes associated with the pathology are given for the most statistically significant pathology.

Table 4. Statistical significance of overrepresentation of biological processes based on gene-regulatory list analysis

Group of Biological Processes (BPs)	Number of BPs in the group	The most significant BP	FDR	Number of genes
Regulation of apoptosis	7	Positive regulation of apoptosis	4.1×10^{-7}	14
Response to stressors	3	Cellular response to mechanical stimulus	4.1×10^{-7}	9
Regulation of cell signaling pathways	9	Positive regulation of MAP kinase activity	4.0×10^{-7}	9
Regulation of transcription	8	Positive regulation of transcription from RNA-pol. II promoter	3.9×10^{-7}	23
Regulation of cardiovascular cell proliferation	7	Positive regulation of angiogenesis	6.8×10^{-5}	9
Regulation of cell proliferation	3	Positive regulation of cell proliferation	0.0014	12
Inflammatory processes	5	Positive regulation of inflammatory response	0.0046	6

Note. False Discovery Rate (FDR) and the number of genes associated with BPs are given for the most statistically significant BPs.

specifically associated with delirium (García-Bueno et al., 2016a, b). Interestingly, the literature discusses the association of preoperative pain factors with depressive symptoms and the subsequent development of POD (O'Sullivan et al., 2014). When considering the regulators of gene expression, among the significant pathologies, the group of pathologies of the cardiovascular system was predominant, which was expected, given the patients' history. In this regard, it can be assumed that a special role in the manifestation of pathological mechanisms of delirium belongs to the regulation of the activity of protein products and, to a lesser extent, to the regulation of gene expression. Note that no significant differences between the two groups of regulators were found as a result of the analysis of BP overrepresentation.

An important structural characteristic of the graph of gene networks, which determines the peculiarities of their functioning, is the centrality of vertices. One of its indicators is the degree centrality of vertices, which characterizes the ratio of the number of links of a given vertex to the total number of links in the graph and is widely used in the analysis of gene networks. The enzyme sphingomyelinase (ASM, see the Figure) had the largest number of connections (regulation of activity, degradation, and transport) with regulatory proteins among the graph vertices corresponding to enzymes. This enzyme cleaves sphingomyelins into phosphatidylcholine and ceramide, which have a signaling function. The function of ASM enzyme was modulated by 10 regulatory proteins, 6 of which had the "activity regulation" type of links (ASM3B, Hsp70, KLRB1, TNFA, TNR6, VEGFA), 3 proteins (CASP8, SORT, TNR5) had the "transport regulation" type, and there was one link with CASP7 protein with the "proteolysis" type. Note that among the regulatory proteins, caspase-8 (CASP8) and tumor necrosis factor alpha (TNFA) were present and found to be associated with overrepresented diseases such as epilepsy, depression, dementia and other neuropsychiatric diseases. According to the literature, CASP8 accomplishes the activation and translocation of ASM to the surface of the plasma membrane. ASM activation results in the cleavage of sphingomyelins and the formation of ceramide, which promotes caspase-8 activity and induction of apoptosis (Grassmé et al., 2003). Surgical interventions are known to provoke the penetration of pro-inflammatory factors such as interleukins and TNFA across the GEB, which contributes to neuroinflammation and may be associated with the development of POD (Alam et al., 2018). According to the reconstructed gene network, TNFA increases phosphomyelinase activity (Corre et al., 2013) and is also associated with overrepresented neuropsychiatric diseases such as depression, epilepsy, etc. (see Supplementary Table S3).

The *SPHK2* gene (see the Figure), encoding the enzyme sphingosine kinase 2, had the highest centrality index among the gene network graph nodes corresponding to the genes. The gene network represented 7 regulators of expression of this gene encoded by the *AGT*, *CCNA1*, *FAS*, *IL17A*, *KCNN1*, *SPHK1*, and *PAPSS1* genes (see the Figure). In contrast to the peak corresponding to the ASM protein, there were no regulators of *SPHK2* expression associated with neuropsychiatric diseases. This fact once again indicates that the most important contribution to the dysfunction

of the sphingolipid metabolism pathway associated with postoperative delirium may come not from the regulation of the expression of genes encoding enzymes of the metabolic pathway, but from the impaired transport, activity, and stability of the products of these genes. Genes associated with other disease groups were represented among the regulators of SPHK2 expression (see Supplementary Table S5). For example, fatty acid synthase (FAS) activity is associated with myocardial infarction, hypertension, type II diabetes, and other diseases (Nosrati-Oskouie et al., 2021).

Conclusion

An integrated approach in metabolomic analysis of blood plasma from cardiac surgery patients using HPLC-MS/MS and bioinformatic methods of ANDSystem gene network reconstruction allowed us to identify potential markers of the sphingomyelin class, as well as regulatory genes, the dysfunction of which may underlie the mechanisms of postoperative delirium (POD) development. The analysis of disease overrepresentation revealed that the groups of pathologies such as neuropsychiatric diseases, cardiac and renal pathologies, inflammatory processes, and surgical intervention were associated with these regulatory proteins. The function of regulators associated with CVDs could be impaired in patients with POD due to heart surgery and medical procedures such as artificial circulation (Gao et al., 2005). However, since heart surgery was undergone by all subjects, it can be expected that the altered function of these regulatory proteins could have equally affected both the group with and without POD. In this regard, the function of a group of regulators associated with neuropsychiatric diseases could have been specifically impaired in patients with POD, which was responsible for the decreased plasma sphingolipid content in these patients.

Among the nodes of the gene network graph, the node with 10 regulatory connections corresponding to the ASM enzyme (phosphomyelinase) had the highest centrality index. Proteins encoded by the *TNFA*, *CASP8*, *TNR5*, and *VEGFA* genes, which are associated with epilepsy, depression, and other neuropsychiatric diseases, were found among regulators of ASM activity and transport. Among the nodes corresponding to the genes, the *SPHK2* (sphingosine kinase 2) gene had the highest centrality score in the graph. The expression of this gene is regulated by 7 proteins encoded by the *AGT*, *CCNA1*, *FAS*, *IL17A*, *KCNN1*, *SPHK1*, and *PAPSS1* genes.

The proposed hypotheses on the role of regulatory genes in the development of AMP can be used to plan transcriptomic and proteomic analysis experiments to study the molecular genetic mechanisms of this complication.

References

Alam A., Hana Z., Jin Z., Suen K.C., Ma D. Surgery, neuroinflammation and cognitive impairment. *EBioMedicine*. 2018;37:547-556. DOI 10.1016/j.ebiom.2018.10.021

Basov N.V., Rogachev A.D., Aleshkova M.A., Gaisler E.V., Sotnikova Y.S., Patrushev Y.V., Tolstikova T.G., Yarovaya O.I., Pokrovsky A.G., Salakhutdinov N.F. Global LC-MS/MS targeted metabolomics using a combination of HILIC and RP LC separation modes on an organic monolithic column based on 1-vinyl-1,2,4-triazole. *Talanta*. 2024;267:125168. DOI 10.1016/j.talanta.2023.125168

Brown C.H. Delirium in the cardiac surgical intensive care unit. *Curr. Opin. Anaesthesiol*. 2014;27(2):117-122. DOI 10.1097/ACO.0000000000000061

Cerejeira J., Firmino H., Vaz-Serra A., Mukaetova-Ladinska E.B. The neuroinflammatory hypothesis of delirium. *Acta Neuropathol*. 2010; 119(6):737-775. DOI 10.1007/s00401-010-0674-1

Corre I., Guillonneau M., Paris F. Membrane signaling induced by high doses of ionizing radiation in the endothelial compartment. Relevance in radiation toxicity. *Int. J. Mol. Sci*. 2013;14(11):22678-22696. DOI 10.3390/ijms141122678

Gao L., Taha R., Gauvin D., Othmen L.B., Wang Y., Blaise G. Postoperative cognitive dysfunction after cardiac surgery. *Chest*. 2005; 128(5):3664-3670. DOI 10.1378/chest.128.5.3664

García-Bueno B., Gassó P., MacDowell K.S., Callado L.F., Mas S., Bernardo M., Lafuente A., Meana J.J., Leza J.C. Evidence of activation of the Toll-like receptor-4 proinflammatory pathway in patients with schizophrenia. *J. Psychiatry Neurosci.* 2016a;41(3):E46-E55. DOI 10.1503/jpn.150195

García Bueno B., Caso J.R., Madrigal J.L., Leza J.C. Innate immune receptor Toll-like receptor 4 signalling in neuropsychiatric diseases. *Neurosci. Biobehav. Rev.* 2016b;64:134-147. DOI 10.1016/j.neubiorev.2016.02.013

Gottesman R.F., Grega M.A., Bailey M.M., Pham L.D., Zeger S.L., Baumgartner W.A., Selnes O.A., McKhann G.M. Delirium after coronary artery bypass graft surgery and late mortality. *Ann. Neurol.* 2010;67(3):338-344. DOI 10.1002/ana.21899

Grassmé H., Cremesti A., Kolesnick R., Gulbins E. Ceramide-mediated clustering is required for CD95-DISC formation. *Oncogene*. 2003; 22(35):5457-5470. DOI 10.1038/sj.onc.1206540

Guo Y., Li Y., Zhang Y., Fang S., Xu X., Zhao A., Zhang J., Li J.V., Ma D., Jia W., Jiang W. Post-operative delirium associated with metabolic alterations following hemi-arthroplasty in older patients. *Age Ageing*. 2019;49(1):88-95. DOI 10.1093/ageing/afz132

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*. 2009;4(1):44-57. DOI 10.1038/nprot.2008.211

Huang H., Han J., Li Y., Yang Y., Shen J., Fu Q., Chen Y. Early serum metabolism profile of post-operative delirium in elderly patients following cardiac surgery with cardiopulmonary bypass. *Front. Aging Neurosci*. 2022;14:857902. DOI 10.3389/fnagi.2022.857902

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(Suppl.11):228. DOI 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved ai-based short names recognition. *Int. J. Mol. Sci*. 2022;23(23):14934. DOI 10.3390/ijms232314934

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Sys. Biol.* 2015;9(Suppl.2):S2. DOI 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019; 20(Suppl.1):34. DOI 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci. Rep*. 2022;12(1):19977. DOI 10.1038/s41598-022-24170-0

Kanehisa M. The KEGG Database. In: 'In silico' Simulation of Biological Processes: Novartis Foundation Symposium. Chichester, UK: John Wiley & Sons, 2002;247:91-103. DOI 10.1002/0470857897.ch8

Kanehisa M., Sato Y., Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci*. 2022;31(1): 47-53. DOI 10.1002/pro.4172

Ke C., Pan C.W., Zhang Y., Zhu X., Zhang Y. Metabolomics facilitates the discovery of metabolic biomarkers and pathways for ischemic stroke: a systematic review. *Metabolomics*. 2019;15(12):152. DOI 10.1007/s11306-019-1615-1

Li K., Naviaux J.C., Bright A.T., Wang L., Naviaux R.K. A robust, single-injection method for targeted, broad-spectrum plasma metabolomics. *Metabolomics*. 2017;13(10):122. DOI 10.1007/s11306-017-1264-1

Morimoto Y., Yoshimura M., Utada K., Setoyama K., Matsumoto M., Sakabe T. Prediction of postoperative delirium after abdominal surgery in the elderly. *J. Anesth*. 2009;23(1):51-56. DOI 10.1007/s00540-008-0688-1

Nosrati-Oskouie M., Aghili-Moghaddam N.S., Sathyapalan T., Sahebkar A. Impact of curcumin on fatty acid metabolism. *Phytother. Res*. 2021;35(9):4748-4762. DOI 10.1002/ptr.7105

O'Sullivan R., Inouye S.K., Meagher D. Delirium and depression: inter-relationship and clinical overlap in elderly people. *Lancet Psychiatry*. 2014;1(4):303-311. DOI 10.1016/S2215-0366(14)70281-0

Patrushev Y.V., Sotnikova Y.S., Sidel'nikov V.N. A monolithic column with a sorbent based on 1-vinyl-1,2,4-triazole for hydrophilic HPLC. *Protect. Met. Phys. Chem. Surf*. 2020;56(1):49-53. DOI 10.1134/s2070205119060248

Pisani M.A., Kong S.Y.J., Kasl S.V., Murphy T.E., Araujo K.L.B., Ness P.H.V. Days of delirium are associated with 1-year mortality in an older intensive care unit population. *Am. J. Resp. Crit. Care Med*. 2009;180(11):1092-1097. DOI 10.1164/rccm.200904-0537OC

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. DOI 10.3390/metabo11030133

Squiccimarro E., Labriola C., Malvindi P.G., Margari V., Guida P., Visicchio G., Kounakis G., Favale A., Dambruoso P., Mastrototaro G., Lorusso R., Paparella D. Prevalence and clinical impact of systemic inflammatory reaction after cardiac surgery. *J. Cardiothorac. Vasc. Anesth*. 2019;33(6):1682-1690. DOI 10.1053/j.jvca.2019.01.043

Stafford-Smith M., Patel U.D., Phillips-Bute B.G., Shaw A.D., Swaminathan M. Acute kidney injury and chronic kidney disease after cardiac surgery. *Adv. Chronic Kidney Dis*. 2008;15(3):257-277. DOI 10.1053/j.ackd.2008.04.006

Steiner L.A. Postoperative delirium. Part 1: Pathophysiology and risk factors. *Eur. J. Anaesthesiol*. 2011;28(9):628-636. DOI 10.1097/EJA.0b013e328349b7f5

Tripp B.A., Dillon S.T., Yuan M., Asara J.M., Vasunilashorn S.M., Fong T.G., Metzger E.D., Inouye S.K., Xie Z., Ngo L.H., Marcantonio E.R., Libermann T.A., Otu H.H. Targeted metabolomics analysis of postoperative delirium. *Sci. Rep*. 2021;11(1):1521. DOI 10.1038/s41598-020-80412-z

Vutskits L., Xie Z. Lasting impact of general anaesthesia on the brain: mechanisms and relevance. *Nat. Rev. Neurosci*. 2016;17:705-717. DOI 10.1038/nrn.2016.128

Wang Y., Shen X. Postoperative delirium in the elderly: the potential neuropathogenesis. *Aging Clin. Experim. Res*. 2018;30(11):1287-1295. DOI 10.1007/s40520-018-1008-8

Wimmer-Greinecker G., Matheis G., Brieden M., Dietrich M., Oremek G., Westphal K., Winkelmann B.R., Moritz A. Neuropsychological changes after cardiopulmonary bypass for coronary artery bypass grafting. *Thorac. Cardiovasc. Surg*. 1998;46(4):207-212. DOI 10.1055/s-2007-1010226

Xiao M.Z., Liu C.X., Zhou L.G., Yang Y., Wang Y. Postoperative delirium, neuroinflammation, and influencing factors of postoperative delirium: a review. *Medicine*. 2023;102(8):e32991-e32991. DOI 10.1097/MD.0000000000032991

ORCID ID

V.A. Ivanisenko orcid.org/0000-0002-1859-4631
N.V. Basov orcid.org/0000-0001-6390-5796
A.A. Makarova orcid.org/0009-0005-1844-7921
A.S. Venzel orcid.org/0000-0002-7419-5168
A.D. Rogachev orcid.org/0000-0002-3338-8529
P.S. Demenkov orcid.org/0000-0001-9433-8341
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
M.A. Kleshchev orcid.org/0000-0002-7537-2525
G.B. Moroz orcid.org/0000-0002-0154-4662
Y.S. Sotnikova orcid.org/0000-0002-0545-703X
Y.V. Patrushev orcid.org/0000-0002-2078-5488
V.V. Lomivorotov orcid.org/0000-0001-8591-6461
N.A. Kolchanov orcid.org/0000-0001-6800-8787
A.G. Pokrovsky orcid.org/0000-0001-5982-8580

Funding. The study was supported by Russian Science Foundation (Moscow, Russia) grant No. 22-23-01068.

Acknowledgements. The authors express their gratitude to the Center for Collective Use (CCU) "Bioinformatics" for the computational resources and their software, created within the framework of the budget project FWNR-2022-0020. The authors are thankful to Arman Azari, M.D. for proofreading the manuscript.

Conflict of interest. The authors declare no conflict of interest.

Received July 21, 2023. Revised August 12, 2023. Accepted August 24, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):776-783
DOI 10.18699/VJGB-23-90

Original Russian text https://vavilovj-icg.ru/

Molecular-genetic pathways of hepatitis C virus regulation of the expression of cellular factors PREB and PLA2G4C, which play an important role in virus replication

E.L. Mishchenko[1, 2], A.A. Makarova[1], E.A. Antropova[1], A.S. Venzel[1, 2], T.V. Ivanisenko[1, 2], P.S. Demenkov[1, 2, 3], V.A. Ivanisenko[1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
salix@bionet.nsc.ru

Abstract. The participants of Hepatitis C virus (HCV) replication are both viral and host proteins. Therapeutic approaches based on activity inhibition of viral non-structural proteins NS3, NS5A, and NS5B are undergoing clinical trials. However, rapid mutation processes in the viral genome and acquisition of drug resistance to the existing drugs remain the main obstacles to fighting HCV. Identifying the host factors, exploring their role in HCV RNA replication, and studying viral effects on their expression is essential for understanding the mechanisms of viral replication and developing novel, effective curative approaches. It is known that the host factors *PREB* (prolactin regulatory element binding) and *PLA2G4C* (cytosolic phospholipase A2 gamma) are important for the functioning of the viral replicase complex and the formation of the platforms of HCV genome replication. The expression of *PREB* and *PLA2G4C* was significantly elevated in the presence of the HCV genome. However, the mechanisms of its regulation by HCV remain unknown. In this paper, using a text-mining technology provided by ANDSystem, we reconstructed and analyzed gene networks describing regulatory effects on the expression of *PREB* and *PLA2G4C* by HCV proteins. On the basis of the gene network analysis performed, we put forward hypotheses about the modulation of the host factors functions resulting from protein-protein interaction with HCV proteins. Among the viral proteins, NS3 showed the greatest number of regulatory linkages. We assumed that NS3 could inhibit the function of host transcription factor (TF) NOTCH1 by protein-protein interaction, leading to upregulation of *PREB* and *PLA2G4C*. Analysis of the gene networks and data on differential gene expression in HCV-infected cells allowed us to hypothesize further how HCV could regulate the expression of TFs, the binding sites of which are localized within *PREB* and *PLA2G4C* gene regions. The results obtained can be used for planning studies of the molecular-genetic mechanisms of viral-host interaction and searching for potential targets for anti-HCV therapy.
Key words: hepatitis C virus; HCV gene replication; replicase HCV; host factors; gene networks; phospholipase PLA2G4C; PREB protein.

For citation: Mishchenko E.L., Makarova A.A., Antropova E.A., Venzel A.S., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Molecular-genetic pathways of hepatitis C virus regulation of the expression of cellular factors PREB and PLA2G4C, which play an important role in virus replication. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):776-783. DOI 10.18699/VJGB-23-90

Молекулярно-генетические пути регуляции вирусом гепатита С экспрессии клеточных факторов PREB и PLA2G4C, играющих важную роль для репликации вируса

Е.Л. Мищенко[1, 2], А.А. Макарова[1], Е.А. Антропова[1], А.С. Вензель[1, 2], Т.В. Иванисенко[1, 2], П.С. Деменков[1, 2, 3], В.А. Иванисенко[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
salix@bionet.nsc.ru

Аннотация. В репликации генома вируса гепатита С (ВГС) участвуют как вирусные, так и хозяйские белки. Терапевтические подходы, основанные на подавлении активности неструктурных вирусных белков NS3, NS5A, NS5B, проходят клинические испытания разных уровней. Однако быстрые мутационные процессы вирусного генома и приобретение лекарственной устойчивости остаются одними из главных препятствий в борьбе с ВГС.

Идентификация и исследование клеточных факторов, участвующих в репликации РНК ВГС, а также регуляция вирусом их экспрессии важны для понимания механизмов репликации вируса и разработки эффективных подходов противовирусной терапии. Известно, что белок PREB, связывающий регуляторный элемент пролактина, и цитозольная фосфолипаза А2 гамма (PLA2G4C) играют важную роль в формировании платформ репликации РНК ВГС, а также в функционировании вирусной репликазы. Экспрессия генов *PREB* и *PLA2G4C* значительно увеличена в присутствии ВГС, но механизмы ее регуляции вирусными белками до сих пор не изучены. В данной работе с применением технологии текст-майнинга, реализованной в программно-информационной системе ANDSystem, реконструированы генные сети регуляции экспрессии генов человека *PREB* и *PLA2G4C* белками ВГС. На основании анализа генных сетей мы выдвинули гипотезы о регуляторных эффектах белков ВГС на функции хозяйских факторов в результате белок-белковых взаимодействий. Среди вирусных белков наибольшее количество регуляторных связей выявлено у вирусной протеазы NS3. Предположительно NS3 в результате белок-белкового взаимодействия подавляет активность транскрипционного фактора NOTCH1, что обусловливает активацию экспрессии *PREB* и *PLA2G4C*. Анализ генных сетей и данных о дифференциальной экспрессии генов в присутствии ВГС позволил нам также выдвинуть гипотезы о регуляции вирусом экспрессии транскрипционных факторов, сайты связывания которых находятся в районах генов *PREB* и *PLA2G4C*, и действии этих транскрипционных факторов на регуляцию транскрипции *PREB* и *PLA2G4C*. Полученные результаты могут быть использованы при планировании исследований по изучению молекулярно-генетических механизмов взаимодействия вирус–хозяин и поиска потенциальных мишеней для разработки лекарств против ВГС.
Ключевые слова: вирус гепатита С; репликация генома ВГС; репликаза ВГС; хозяйские факторы; генные сети; фосфолипаза PLA2G4C; белок PREB.

Introduction

The Hepatitis C virus (HCV) causes a dangerous liver disease, which, starting asymptomatic, turns into a chronic form and can lead to cirrhosis and hepatocellular carcinoma (Yamane et al., 2013). The HCV genome is represented by a plus-chain RNA (~9,600 nucleotides), encoding structural (Core, E1, E2) and non-structural (p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B) proteins. It also contains 5′- and 3′-untranslated regions (UTR) necessary for translating the viral polyprotein and replicating the viral genome (Bartenschlager et al., 2013). Structural glycoproteins E1 and E2 are localized on the viral bilayer lipid envelope surrounding the nucleocapsid, which consists of multiple copies of the Core protein and RNA genome. The p7 protein has membrane cation channel properties; proteins NS2 and NS3/NS4A are proteases that process the viral polyprotein. NS3 also has helicase activity; NS4B and NS5A can modify endoplasmic reticulum (ER) membranes to form vesicular membrane structures – platforms for the replication of the HCV genome. NS5B is an RNA-dependent RNA polymerase. The complex of non-structural proteins NS3–NS5B, which also involves host factors, performs the function of viral replicase in the host cell (Moradpour et al., 2007). The virus genome is highly heterogeneous due to the high error rate of the RNA-dependent RNA polymerase NS5B. This property of NS5B is considered the main reason for the virus's rapid acquisition of drug resistance (Powdrill et al., 2011).

Currently, a great deal of research is directed towards identifying and studying the properties of cellular factors involved in modifying ER membranes to form vesicle clusters in which the HCV RNA genome replicates, which are part of the viral replicase. For instance, it has been established that the receptor for activated C kinase 1 (RACK1) associates with NS5A and the ATG14L-Beclin1-Vps34-Vps15 autophagosome formation initiation complex, stimulating the formation of vesicular membrane structures (Lee et al., 2019). The early endosome (EE) protein Rab5, regulating endocytosis and EE fusion, and the late endosome (LE) protein Rab7, enhancing LE transport to lysosomes, are associated with NS4B and involved in the biogenesis of these membrane structures (Manna et al., 2010). The small GTPase Rab18·GTP on lipid droplet (LD) membranes interacts with the viral protein NS5A on the ER membrane. The association of LD and ER membranes due to the direct interaction of Rab18·GTP and NS5A leads to the localization of HCV replicase complexes near LDs and stimulates HCV RNA replication (Salloum et al., 2013).

Phosphatidylinositol 4-kinase IIIα (PI4KIIIα) is important in forming membrane vesicular structures and replication complexes. Through protein-protein interaction, NS5A stimulates the activity of PI4KIIIα, leading to the formation of phosphatidylinositol-4-phosphate (PI4P), which recruits and coordinates viral and host proteins on the membrane that contains PI4P-affine lipid-binding domains (Berger et al., 2011; Reiss et al., 2011). Moreover, HCV can regulate the expression of cellular factors that play an important role in virus replication. For example, cytosolic phospholipase A2 gamma (PLA2G4C), which hydrolyzes membrane phosphoglycerides to form free fatty acids and lysophosphatidate and directly affects the structure, shape, merger, and interaction of the membranes with proteins (Brown et al., 2003), has several times increased expression at both RNA and protein levels in the presence of HCV RNA (Xu et al., 2012).

The expression of the *PREB* gene (prolactin regulatory element binding protein) is also significantly increased in the presence of HCV (Kong et al., 2016). The PREB protein functions as a regulatory factor for COPII vesicle budding from the ER membrane (LaPointe et al., 2004), associates with NS4B, is involved in the formation of membrane vesicular structures and is localized in the active HCV replication complex through interaction with NS4B (Kong et al., 2016). Despite accumulated evidence of increased *PREB* and *PLA2G4C* expression in the presence of HCV, the molecular mechanisms regulating the expression of these host factors are poorly understood.

The technology of text mining is a useful tool for studying molecular-genetic interactions. We previously developed the software and information system ANDSystem (Ivanisen-

ko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022), which implements a full cycle of knowledge engineering, including automatic extraction of information from scientific publications and factographic databases, integration, and representation of information in the form of semantic networks in the knowledge base, as well as providing user access to the knowledge base for the reconstruction and analysis of gene networks. ANDSystem was used to solve a wide range of tasks, including analyzing the interactome of Hepatitis C virus proteins with human proteins, interpreting metabolomic analysis results, gene prioritization tasks, searching for new potential drug targets, and others. In particular, the analysis of protein-protein interactions of HCV and human proteins allowed us to reconstruct potential pathways of regulating the external pathway of apoptosis by viral proteins (Saik et al., 2016), as well as to study the features of HCV protein regulation of genes prone to aberrant methylation in hepatocellular carcinoma (Antropova et al., 2022). Based on the data of metabolomic analysis of the blood plasma of patients with COVID-19, regulatory pathways describing the control of human metabolic pathways by SARS-Cov-2 proteins were reconstructed, and it was shown that a number of non-structural viral proteins had the most significant regulatory impact (Ivanisenko V.A. et al., 2022). With the help of reconstruction and analysis of gene networks, new methods of gene prioritization were proposed, which were used to search for candidate genes associated with lymphedema as well as with major depressive disorder (Yankina et al., 2018; Saik et al., 2019). Using ANDSystem, new potential pharmacological targets for treating comorbid conditions of asthma and hypertension were proposed (Saik et al., 2018a, b).

In our work, using the ANDSystem software information system, we reconstructed and analyzed the pathways of HCV protein regulation of the expression of cellular factor genes *PLA2G4C* and *PREB*, which play an important role in the formation of membrane vesicular structures – the platform for viral RNA replication, and in the functioning of the viral replicase. Through computer analysis, 28 human transcription factors (TFs) under the control of HCV were found which could participate in the regulation of *PLA2G4C* and *PREB* expression. It turned out that out of these TFs, 16 proteins participate in the regulation of *PLA2G4C*, 23 – in the regulation of *PREB*, and 11 are common. Based on the analysis of gene networks and data on differential gene expression, hypotheses have been put forward about the regulatory effects of viral proteins on the functions of TFs with which they form complexes as a result of protein-protein interactions, as well as the regulatory effects of these TFs on the expression of *PLA2G4C* and *PREB*.

Materials and methods

Obtaining the list of differentially expressed genes (DEGs) of human proteins in the presence of HCV proteins. Using RNA sequencing results available at the NCBI GEO resource (http://www.ncbi.nlm.nih.gov/geo) (Edgar et al., 2002), a list of human genes differentially expressed in Huh7.5.1 hepatocytes under HCV infection conditions was obtained via the GSE66842 identifier. The RNA sequencing results were analyzed using the GEO2R tool, allowing to obtain statistical processing results and data visualization on differential gene expression under experimental conditions. We selected statistically significant DEGs at the control point "10 days after HCV infection" (GSE66842). The study also used transcriptome analysis results of differential gene expression in Huh.7.5 hepatocytes at the control point "72 hours after HCV infection" (Papic et al., 2012). These results were combined into a final list of DEGs to reconstruct gene networks.

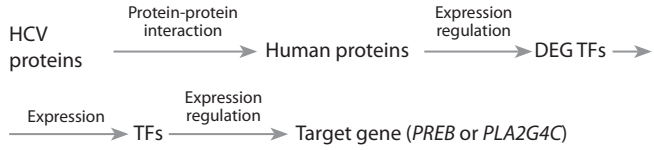


Fig. 1. Scheme for constructing regulatory molecular-genetic pathways for modulating the expression of host factor genes by HCV proteins.

Identification of transcription factors. Transcription factors, the binding sites of which are located in the *PREB* and *PLA2G4C* genes, as well as in flanking regions of these genes within a range of ±2,000 bp, were extracted from the GTRD database (http://gtrd20-06.biouml.org/) (Yevshin et al., 2017; Kolmykov et al., 2021), which integrates studies on genome organization. For gene network construction, the TF genes differentially expressed under Hepatitis C virus infection conditions were selected.

Reconstruction and analysis of molecular genetic pathways of *PREB* and *PLA2G4C* gene expression regulation by HCV proteins using ANDSystem. Molecular genetic pathways for regulating host factors PREB and PLA2G4C expression by HCV proteins were reconstructed using ANDSystem and its graphical user interface ANDVisio. The ANDVisio program accesses the ANDSystem knowledge base, which contains over 40 million facts about intermolecular interactions, including protein-protein interactions, gene expression regulation, activity regulation, degradation, and protein transport.

The construction of regulatory molecular genetic pathways describing interactions between HCV proteins and human proteins and genes was carried out using the "Pathway Master" module of the ANDVisio program. The relationships between the participants of these pathways, including protein-protein interactions and gene expression regulation, are arranged according to the scheme (Fig. 1).

Results and discussion

Reconstruction of the interactome of human proteins and HCV proteins

Using the ANDSystem software and information system, an interactome of 10 HCV proteins with 333 human proteins was reconstructed (Fig. 2). It turned out that 195 human proteins interact with NS3, 59 – with NS5A, 50 – with Core, 26 – with NS5B, 15 – with NS2, 7 – with E2 and p7, 6 – with NS4A, 5 – with E1, 4 proteins – with NS4B. The gene network illustrates that only a few human proteins interact with more than one HCV protein. Among them are transcription factors potentially regulating the expression of target genes *PREB* and *PLA2G4C*.

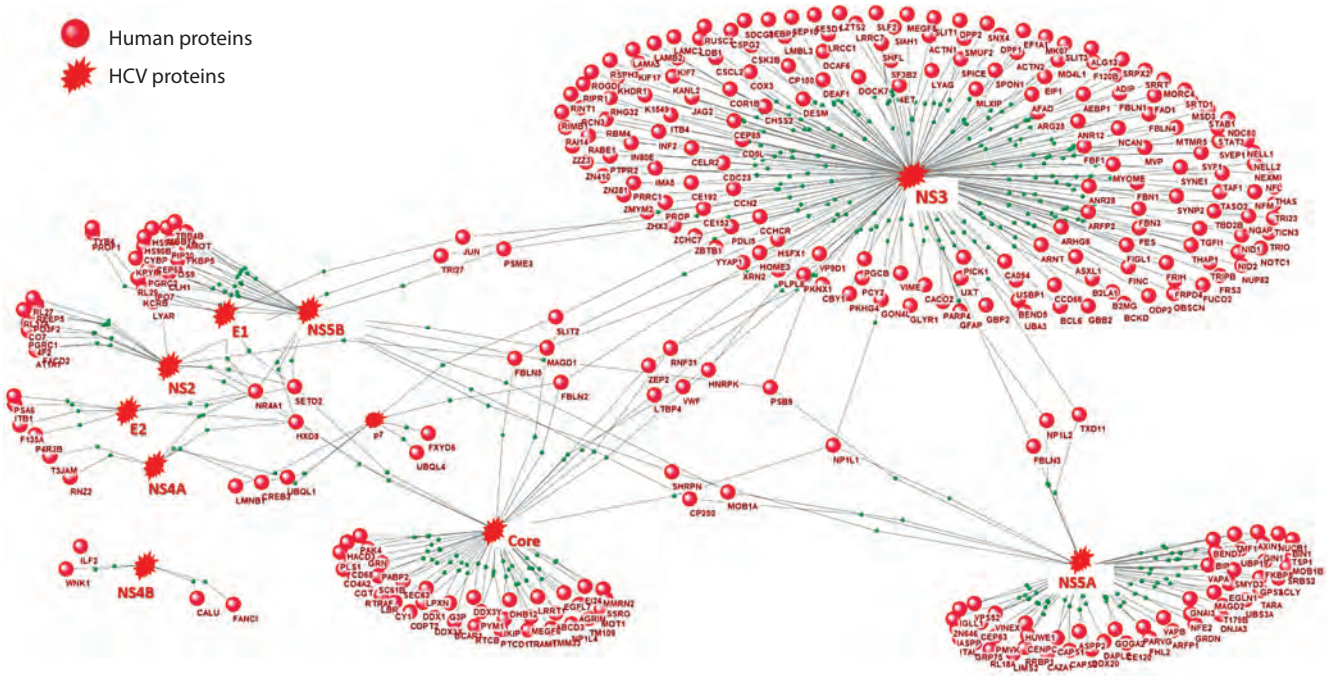


Fig. 2. Graph of interactions between human proteins and HCV proteins, reconstructed using the ANDSystem software and information system. Black lines indicate protein-protein interactions.

Reconstruction of molecular-genetic pathways regulating the expression of *PREB* and *PLA2G4C* genes by HCV proteins

Published scientific results indicate that the expression of cellular factors PLA2G4C (Xu et al., 2012) and PREB (Kong et al., 2016) is significantly enhanced in the presence of HCV proteins. These host factors play an important role in HCV replication. They are involved in forming membranous vesicular structures – compartments of viral RNA replication, and in the functioning of the HCV replicase complex (Xu et al., 2012; Kong et al., 2016). However, the molecular-genetic mechanisms for increasing the expression of *PREB* and *PLA2G4C* in the context of HCV infection have not been studied to date. Transcription factors (TFs) regulated by viral proteins were identified using information on differential gene expression. It should be noted that in our study, we did not consider TFs, the expression of which did not change under conditions of HCV infection. The GTRD database extracted lists containing 432 and 693 TFs, the binding sites of which are in the regions of *PREB* and *PLA2G4C* genes, respectively. Among many transcription factors, 92 TFs were selected, the genes of which are differentially expressed in the presence of HCV proteins (69 and 63 TF genes for *PREB* and *PLA2G4C*, respectively, and 40 TFs common for both target genes).

Using ANDSystem, the molecular-genetic pathways regulating the expression of *PREB* and *PLA2G4C* by HCV proteins were reconstructed and analyzed (Figs. 3 and 4). Among the regulatory pathways, the first layer of which were HCV proteins, and the final ones were *PREB* and *PLA2G4C* genes, there turned out to be 28 out of 92 TFs, indicating the regulation of these TFs by viral proteins.

Figure 3 illustrates the regulatory molecular-genetic pathways of *PREB* expression by HCV proteins. These pathways include 24 proteins presented in layer 2, 23 participants in layer 4, and their encoding genes in layer 3. As shown in the gene network graph, only 23 out of 69 TFs were included in the regulatory pathways, suggesting that these specific TFs may regulate the transcription of the *PREB* gene under HCV infection conditions.

The gene network in Figure 4 illustrates the pathways of *PLA2G4C* expression regulation by HCV proteins. In the GTRD database, 63 TF binding sites were found in the regulatory regions of the *PLA2G4C* gene, which are differentially expressed genes (DEGs). Only 16 out of these 63 TFs were part of the regulatory pathways. This suggests that these particular TFs presumably regulate the transcription of the *PLA2G4C* gene under HCV infection conditions. It was previously shown that the NS3 protein of the Hepatitis C virus stimulates the activity of the TF STAT3 (Machida et al., 2006). Moreover, STAT3 significantly enhances the transcription of the *MYC* gene (Kiuchi et al., 1999; Papic et al., 2012). Furthermore, it was demonstrated in a study (Xiong et al., 2017) that the alteration of MYC expression enhanced *PLA2G4C* expression, which aligns with the regulatory pathway we identified. Similarly, the positive regulation of *XBP1* expression by STAT3 (Diehl et al., 2008) and the increased expression of *XBP1* (Papic et al., 2012) in the presence of HCV may account for the activating effect of *XBP1* on *PLA2G4C* transcription.

The use of ANDSystem allowed us to propose hypotheses about the regulation of TF expression by HCV proteins interacting with the regulatory region sites of the *PREB* and *PLA2G4C* genes (see Figs. 3 and 4). It should be noted that 11 TFs were simultaneously represented among the regulators of *PREB* and *PLA2G4C*. Based on the data on differential gene expression and the nature of regulatory molecular-genetic pathway connections, we can hypothesize about the effect

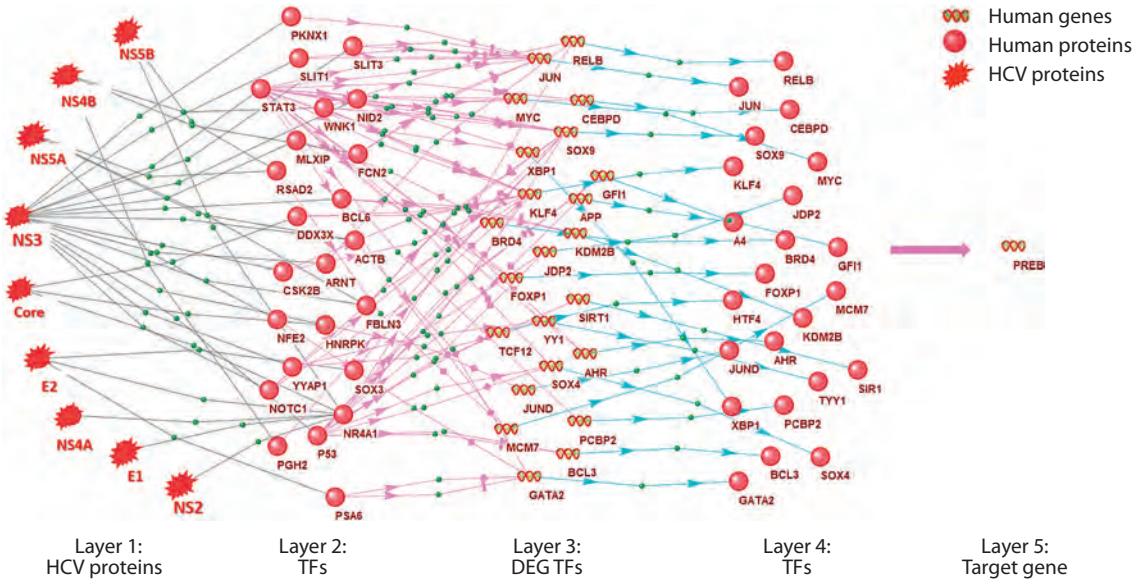


Fig. 3. Gene network of molecular-genetic pathways regulating the expression of the *PREB* gene in conditions of HCV infection.
Here and in Fig. 4: black lines – protein-protein interactions; pink arrows – expression regulation; blue arrows – expression.

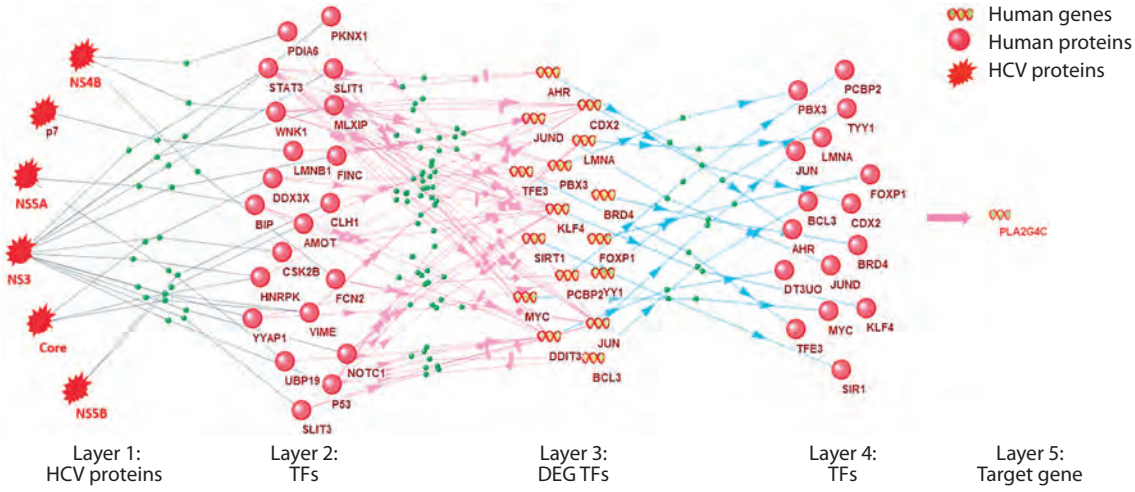


Fig. 4. Gene network of molecular-genetic pathways regulating the expression of the *PLA2G4C* gene under HCV infection conditions.

these TFs (layer 4) have on the transcription of *PREB* and *PLA2G4C* (see the Table). For example, the increased expression of the layer 3 TF gene and positive regulation by the layer 2 TF may lead to the activation of *PREB* and *PLA2G4C* transcription. Specifically, from the regulatory pathways, it follows that the TF CEBPD positively regulates the expression of *PREB*, as the expression of *CEBPD* is positively regulated by STAT3 (layer 2) and is elevated in the presence of HCV (Papic et al., 2012). Conversely, the reduced expression of the layer 4 TF in the presence of HCV and the negative sign of expression regulation between layer 2 and 3 participants explain the inhibitory effect of the TF on the transcription of *PREB* and *PLA2G4C*.

The studies show that the Core HCV protein increases the expression of *NR4A1* (Tan, Li, 2015), while the transcription factor NR4A1 inhibits the expression of the *SOX9* gene (Hu et al., 2014). In the regulatory pathways we reconstructed, NR4A1 is a transcription factor of layer 2, interacts with six HCV proteins (Core, E1, E2, NS2, NS4A, NS5B), and has a negative effect on *SOX9*. Therefore, the transcription factor SOX9, inhibited at the RNA level under HCV infection conditions, presumably reduces the expression of the *PREB* gene. The hypotheses we proposed based on gene network analysis should be experimentally confirmed in the future.

Analyzing the reconstructed gene networks allowed us to propose hypotheses about how viral proteins might affect

The expected effect of layer 4 TFs on the expression of *PREB* and *PLA2G4C*

TF	Expected effect*		TF	Expected effect		TF	Expected effect	
	PREB	*PLA2G4C*		*PREB*	*PLA2G4C*		*PREB*	*PLA2G4C*
AHR	↑	↑	JDP2	↑	–	RELB	↑	–
APP	↑	–	JUN	↑	↑	SIRT1	↑	↑
BCL3	↑	↑	JUND	↑	↑	SOX4	↑	–
BRD4	↑	↑	KDM2B	↑	–	SOX9	↓	–
CDX2	–	↑	KLF4	↑	↑	TCF12	↑	–
CEBPD	↑	–	LMNA	–	↑	TFE3	–	↑
DDIT3	↑	↑	MCM7	↑	–	XBP1	↑	–
FOXP1	↓	↓	MYC	↑	↑	YY1	↑	↑
GATA2	↑	–	PBX3	–	↑			
GFI1	↑	–	PCBP2	↑	↑			

* «↑» – positive regulation, «↓» – negative regulation, «–» – no regulation.

the function of TFs with which they form complexes due to protein-protein interactions. These hypotheses were based on the structure of regulatory molecular-genetic pathways and data on differential gene expression, similar to the hypotheses about regulating *PREB* and *PLA2G4C* by TFs. A viral protein has a negative effect on the function of a protein from layer 2 of the regulatory pathway as a result of physical interaction with it in the following cases: (1) the layer 2 participant is connected to a participant from layer 3 by positive regulation of expression type, and the expression of the layer 3 participant is reduced in the presence of HCV; (2) the layer 2 participant is connected to a participant from layer 3 by negative regulation of expression type, and the expression of the layer 3 participant is increased in the presence of HCV. A viral protein has a positive effect on the function of a protein from layer 2 in the following cases: (1) the layer 2 participant is connected to a participant from layer 3 by positive regulation of expression type, and the expression of the layer 3 participant is increased in the presence of HCV; (2) the layer 2 participant is connected to a participant from layer 3 by negative regulation of expression type, and the expression of the layer 3 participant is reduced in the presence of HCV.

According to the reconstructed regulatory molecular-genetic pathways, the largest number of regulatory connections among the HCV proteins was identified for the viral protease NS3. One of the proteins directly interacting with NS3 is the TF NOTCH1. Numerous scientific studies of this TF have been published; however, we did not find information about the effect of NS3 on the function of NOTCH1 due to protein-protein interactions. From analyzing regulatory pathways and differential gene expression data, we hypothesized that NS3 suppresses NOTCH1 activity due to protein-protein interaction. It was previously shown that NOTCH1 activates the transcription of *SOX9* (Zong et al., 2009) and inhibits *KLF4* (Xue et al., 2016), which would lead to a negative effect on the transcription of *PREB* and *PLA2G4C*. However, the actual change in the expression of target genes and their TFs SOX9 and KLF4 aligns with the hypothesis about the suppression of NOTCH1 activity by the viral protein NS3.

Conclusion

Using the ANDSystem software system, molecular-genetic pathways of regulation of *PLA2G4C* and *PREB* gene expression by Hepatitis C virus proteins have been reconstructed and analyzed. The protein products of these genes are essential for HCV replication, as they participate in the modification of membranes with the formation of membrane vesicle clusters, which are compartments of HCV genome replication and are also involved in the composition and functioning of the HCV replicase. The theoretical data obtained in our work can be useful for planning studies on the mechanisms by which HCV uses human proteins for its genome replication and for searching for potential targets for antiviral therapy.

References

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Lavrik I.N., Ivanisenko V.A. Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. DOI 10.18699/VJGB-22-89

Bartenschlager R., Lohmann V., Penin F. The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection. *Nat. Rev. Microbiol*. 2013;11(7):482-496. DOI 10.1038/nrmicro3046

Berger K.L., Kelly S.M., Jordan T.X., Tartell M.A., Randall G. Hepatitis C virus stimulates the phosphatidylinositol 4-kinase III alpha-dependent phosphatidylinositol 4-phosphate production that is essential for its replication. *J. Virol*. 2011;85(17):8870-8883. DOI 10.1128/JVI.00059-11

Brown W.J., Chambers K., Doody A. Phospholipase A2 (PLA2) enzymes in membrane trafficking: mediators of membrane shape and

function. *Traffic*. 2003;4(4):214-221. DOI 10.1034/j.1600-0854.2003.00078.x

Diehl S.A., Schmidlin H., Nagasawa M., van Haren S.D., Kwakkenbos M.J., Yasuda E., Beaumont T., Scheeren F.A., Spits H. STAT3-mediated up-regulation of BLIMP1 is coordinated with BCL6 down-regulation to control human plasma cell differentiation. *J. Immunol.* 2008;180(7):4805-4815. DOI 10.4049/jimmunol.180.7.4805

Edgar R., Domrachev M., Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. DOI 10.1093/nar/30.1.207

Hu Y.W., Zhang P., Yang J.Y., Huang J.L., Ma X., Li S.F., Zhao J.Y., Hu Y.R., Wang Y.C., Gao J.J., Sha Y.H., Zheng L., Wang Q. Nur77 decreases atherosclerosis progression in apoE$^{-/-}$ mice fed a high-fat/high-cholesterol diet. *PLoS One*. 2014;9(1):e87313. DOI 10.1371/journal.pone.0087313

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(Suppl.11):228. DOI 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int. J. Mol. Sci.* 2022;23(23):14934. DOI 10.3390/ijms232314934

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst Biol*. 2015;9(Suppl.2):S2. DOI 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. DOI 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci. Rep*. 2022;12(1):19977. DOI 10.1038/s41598-022-24170-0

Kiuchi N., Nakajima K., Ichiba M., Fukada T., Narimatsu M., Mizuno K., Hibi M., Hirano T. STAT3 is required for the gp130-mediated full activation of the c-*myc* gene. *J. Exp. Med*. 1999;189(1):63-73. DOI 10.1084/jem.189.1.63

Kolmykov S., Yevshin I., Kulyashov M., Sharipov R., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., Kolpakov F. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res*. 2021;49(D1):D104-D111. DOI 10.1093/nar/gkaa1057

Kong L., Fujimoto A., Nakamura M., Aoyagi H., Matsuda M., Watashi K., Suzuki R., Arita M., Yamagoe S., Dohmae N., Suzuki T., Sakamaki Y., Ichinose S., Suzuki T., Wakita T., Aizaki H. Prolactin regulatory element binding protein is involved in hepatitis C virus replication by interaction with NS4B. *J. Virol*. 2016;90(6):3093-3111. DOI 10.1128/JVI.01540-15

LaPointe P., Gurkan C., Balch W.E. Mise en place – this bud's for the Golgi. *Mol. Cell.* 2004;14(4):413-414. DOI 10.1016/s1097-2765(04)00267-9

Lee J.S., Tabata K., Twu W.-I., Rahman M.S., Kim H.S., Yu J.B., Jee M.H., Bartenschlager R., Jang S.K. RACK1 mediates rewiring of intracellular networks induced by hepatitis C virus infection. *PLoS Pathog*. 2019;15(9):e1008021. DOI 10.1371/journal.ppat.1008021

Machida K., Cheng K.T., Lai C.K., Jeng K.S., Sung V.M., Lai M.M. Hepatitis C virus triggers mitochondrial permeability transition with production of reactive oxygen species, leading to DNA damage and STAT3 activation. *J. Virol*. 2006;80(14):7199-7207. DOI 10.1128/jvi.00321-06

Manna D., Aligo J., Xu C., Park W.S., Koc H., Heo W.D., Konan K.V. Endocytic Rab proteins are required for hepatitis C virus replication complex formation. *Virology*. 2010;398(1):21-37. DOI 10.1016/j.virol.2009.11.034

Moradpour D., Penin F., Rice C.M. Replication of hepatitis C virus. *Nat. Rev. Microbiol*. 2007;5(6):453-463. DOI 10.1038/nrmicro1645

Papic N., Maxwell C.I., Delker D.A., Liu S., Bret S.E., Heale B.S.E., Hagedorn C.H. RNA-sequencing analysis of 5′ capped RNAs identifies many new differentially expressed genes in acute hepatitis C virus infection. *Viruses*. 2012;4(4):581-612. DOI 10.3390/v4040581

Powdrill M.H., Tchesnokov E.P., Kozak R.A., Russell R.S., Martin R., Svarovskaia E.S., Mo H., Kouyos R.D., Gotte M. Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc. Natl. Acad. Sci. USA*. 2011;108(51):20509-20513. DOI 10.1073/pnas.1105797108

Reiss S., Rebhan I., Backes P., Romero-Brey I., Erfle H., Matula P., Kaderali L., Poenisch M., Blankenburg H., Hiet M.S., Longerich T., Diehl S., Ramirez F., Balla T., Rohr K., Kaul A., Buhler S., Pepperkok R., Lengauer T., Albrecht M., Eils R., Schirmacher P., Lohmann V., Bartenschlager R. Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell Host Microbe*. 2011;9(1):32-45. DOI 10.1016/j.chom.2010.12.002

Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res*. 2016;218:40-48. DOI 10.1016/j.virusres.2015.12.003

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Dosenko V.E., Zolotareva O.I., Choynzonov E.L., Hofestaedt R., Ivanisenko V.A. Search for new candidate genes involved in the comorbidity of asthma and hypertension based on automatic analysis of scientific literature. *J. Integr. Bioinform*. 2018a;15(4):20180054. DOI 10.1515/jib-2018-0054

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofestaedt R., Lavrik I.N., Rogaev E.I. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018b;11(1):61-76. DOI 10.1186/s12920-018-0331-4

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med. Genomics*. 2019;12(Suppl.2):117-131. DOI 10.1186/s12920-019-0492-9

Salloum S., Wang H., Ferguson C., Parton R.G., Tai A.W. Rab18 binds to hepatitis C virus NS5A and promotes interaction between sites of viral replication and lipid droplets. *PLoS Pathog*. 2013;9(8):e1003513. DOI 10.1371/journal.ppat.1003513

Tan Y., Li Y. HCV core protein promotes hepatocyte proliferation and chemoresistance by inhibiting NR4A1. *Biochem. Biophys. Res. Commun.* 2015;466(3):592-598. DOI 10.1016/j.bbrc.2015.09.091

Xiong J., Wang L., Fei X.C., Jiang X., Zheng Z., Zhao Y., Wang C., Li B., Chen S., Janin A., Gale R.P., Zhao W. MYC is a positive regulator of choline metabolism and impedes mitophagy-dependent necroptosis in diffuse large B-cell lymphoma. *Blood Cancer J.* 2017;7(7):e582. DOI 10.1038/bcj.2017.61

Xu S., Pei R., Guo M., Han Q., Lai J., Wang Y., Wu C., Zhou Y., Lu M., Chen X. Cytosolic phospholipase A2 gamma is involved in hepatitis C virus replication and assembly. *J. Virol*. 2012;86(23):13025-13037. DOI 10.1128/JVI.01785-12

Xue Y.K., Tan J., Dou D.W., Chen D., Chen L.J., Ren H.P., Chen L.B., Xiong X.G., Zheng H. Effect of Kruppel-like factor 4 on Notch

pathway in hepatic stellate cells. *J. Huazhong Univ. Sci. Technolog. Med. Sci*. 2016;36(6):811-816. DOI 10.1007/s11596-016-1667-7

Yamane D., McGivern D.R., Masaki T., Lemon S.M. Liver injury and disease pathogenesis in chronic hepatitis C. *Curr. Top. Microbiol. Immunol*. 2013;369:263-288. DOI 10.1007/978-3-642-27340-7_11

Yankina M.A., Saik O.V., Ivanisenko V.A., Demenkov P.S., Khusnutdinova E.K. Evaluation of prioritization methods of extrinsic apoptotic signaling pathway genes for retrieval of the new candidates associated with major depressive disorder. *Russ. J. Genet*. 2018; 54(11):1366-1374. DOI 10.1134/S1022795418110170

Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res*. 2017;45(D1):D61-D67. DOI 10.1093/nar/gkw951

Zong Y., Panikkar A., Xu J., Antoniou A., Raynaud P., Lemaigre F., Stanger B.Z. Notch signaling controls liver development by regulating biliary differentiation. *Development*. 2009;136(10):1727-1739. DOI 10.1242/dev.029140

ORCID ID

A.A. Makarova orcid.org/0009-0005-1844-7921
E.A. Antropova orcid.org/0000-0003-2158-3252
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
P.S. Demenkov orcid.org/0000-0001-9433-8341
V.A. Ivanisenko orcid.org/0000-0002-1859-4631

Acknowledgements. This work is supported by the budget project No. FWNR-2022-0020.

Conflict of interest. The authors declare no conflict of interest.

Received July 19, 2023. Revised August 27, 2023. Accepted August 30, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):784-793
DOI 10.18699/VJGB-23-91

Original Russian text https://vavilovj-icg.ru/

Prioritization of potential pharmacological targets for the development of anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway: the reconstruction and analysis of associative gene networks help

P.S. Demenkov[1, 2, 3], E.A. Antropova[1] ✉, A.V. Adamovskaya[1, 3], E.L. Mishchenko[1, 2], T.M. Khlebodarova[1, 2], T.V. Ivanisenko[1, 2, 3], N.V. Ivanisenko[1], A.S. Venzel[1, 2, 3], I.N. Lavrik[4], V.A. Ivanisenko[1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany
✉ nzhenia@bionet.nsc.ru

Abstract. Hepatocellular carcinoma (HCC) is a common severe type of liver cancer characterized by an extremely aggressive course and low survival rates. It is known that disruptions in the regulation of apoptosis activation are some of the key features inherent in most cancer cells, which determines the pharmacological induction of apoptosis as an important strategy for cancer therapy. The computer design of chemical compounds capable of specifically regulating the external signaling pathway of apoptosis induction represents a promising approach for creating new effective ways of therapy for liver cancer and other oncological diseases. However, at present, most of the studies are devoted to pharmacological effects on the internal (mitochondrial) apoptosis pathway. In contrast, the external pathway induced via cell death receptors remains out of focus. Aberrant gene methylation, along with hepatitis C virus (HCV) infection, are important risk factors for the development of hepatocellular carcinoma. The reconstruction of gene networks describing the molecular mechanisms of interaction of aberrantly methylated genes with key participants of the extrinsic apoptosis pathway and their regulation by HCV proteins can provide important information when searching for pharmacological targets. In the present study, 13 criteria were proposed for prioritizing potential pharmacological targets for developing anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway. The criteria are based on indicators of the structural and functional organization of reconstructed gene networks of hepatocarcinoma, the extrinsic apoptosis pathway, and regulatory pathways of virus-extrinsic apoptosis pathway interaction and aberrant gene methylation-extrinsic apoptosis pathway interaction using ANDSystem. The list of the top 100 gene targets ranked according to the prioritization rating was statistically significantly (p-value = 0.0002) enriched for known pharmacological targets approved by the FDA, indicating the correctness of the prioritization method. Among the promising potential pharmacological targets, six highly ranked genes (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4*, and *KHDRBS1*) are likely to deserve close attention.
Key words: gene networks; hepatocarcinoma; programmed cell death; apoptosis; methylation.

For citation: Demenkov P.S., Antropova E.A., Adamovskaya A.V., Mishchenko E.L., Khlebodarova T.M., Ivanisenko T.V., Ivanisenko N.V., Venzel A.S., Lavrik I.N., Ivanisenko V.A. Prioritization of potential pharmacological targets for the development of anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway: the reconstruction and analysis of associative gene networks help. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):784-793. DOI 10.18699/VJGB-23-91

Приоритизация потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза, на основе реконструкции и анализа ассоциативных генных сетей

П.С. Деменков[1, 2, 3], Е.А. Антропова[1] ✉, А.В. Адамовская[1, 3], Е.Л. Мищенко[1, 2], Т.М. Хлебодарова[1, 2], Т.В. Иванисенко[1, 2, 3], Н.В. Иванисенко[1], А.С. Вензель[1, 2, 3], И.Н. Лаврик[4], В.А. Иванисенко[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Медицинский факультет Магдебургского университета им. Отто фон Герике, Магдебург, Германия
✉ nzhenia@bionet.nsc.ru

Аннотация. Гепатоцеллюлярная карцинома (ГЦК) – распространенный тяжелый тип рака печени, характеризующийся крайне агрессивным течением и низкой выживаемостью. Известно, что нарушения регуляции активации апоптоза являются одной из ключевых особенностей, свойственной большинству раковых клеток, что определяет фармакологическую индукцию апоптоза как важную стратегию терапии рака. Компьютерный дизайн химических соединений, способных целевым образом регулировать внешний сигнальный путь индукции апоптоза, представляет перспективный подход для создания новых эффективных средств терапии рака печени и других онкологических заболеваний. Однако в настоящее время большинство исследований посвящено фармакологическим воздействиям на внутренний (митохондриальный) путь апоптоза, тогда как внешний путь, индуцируемый посредством клеточных рецепторов смерти, остается вне поля зрения. Аберрантное метилирование генов наряду с инфекцией вирусом гепатита С считаются важными факторами риска развития ГЦК. Реконструкция генных сетей, описывающих молекулярные механизмы взаимодействия аберрантно метилированных генов с ключевыми участниками внешнего пути апоптоза, а также пути их регуляции белками вируса гепатита С, может дать важную информацию при поиске фармакологических мишеней. В настоящей работе были предложены 13 критериев приоритизации потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза. В основу критериев легли показатели структурно-функциональной организации реконструированных с использованием ANDSystem генных сетей ГЦК, внешнего пути апоптоза и регуляторных путей взаимодействия «вирус – внешний путь апоптоза» и «аберрантное метилирование генов – внешний путь апоптоза». Список наиболее приоритетных 100 генов-мишеней, ранжированных согласно рейтингу приоритизации, оказался статистически значимо (p-value = 0.0002) обогащен известными фармакологическими мишенями, одобренными FDA, что указывает на корректность примененного метода приоритизации. Среди перспективных потенциальных фармакологических мишеней могут быть представлены шесть генов-кандидатов (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4* и *KHDRBS1*), занимающих высокое положение в ранжированном списке согласно результатам приоритизации.
Ключевые слова: генные сети; гепатокарцинома; программируемая клеточная гибель; апоптоз; метилирование.

Introduction

Hepatocellular carcinoma (HCC) is the most common tumor pathology of the liver, accounting for over 90 % of all malignant neoplasms of the liver and intrahepatic bile ducts (Llovet et al., 2018). It is characterized by an extremely aggressive course and low survival rate. Unlike most other types of cancer, there are some documented risk factors for the occurrence of HCC, such as infections caused by hepatitis C and B viruses, alcohol, fatty infiltration of the liver, hepatitis, autoimmune or chronic cholestatic diseases (Forner et al., 2012). Studies in the field of hepatocarcinogenesis have shown the critical role of genetic and epigenetic mechanisms leading to the formation of monoclonal populations of aberrant and dysplastic hepatocytes, which exhibit telomere erosion and re-expression of telomerase, microsatellite instability, and irreversible structural changes in genes and chromosomes (Balogh et al., 2016). The phenotype of malignant hepatocytes may be caused by the disruption of a number of genes that function in various regulatory pathways, resulting in different molecular variants of HCC (Thorgeirsson, Grisham, 2002). This characteristic of the pathology makes the reconstruction and analysis of gene networks describing the molecular mechanisms of the disease relevant.

In cancer therapeutic research, a central issue is suppressing cellular proliferation and the induction of programmed cell death. Apoptosis, one of the known mechanisms of programmed cell death, is divided into intrinsic and extrinsic, depending on the pathway of signal induction. The apoptosis signal induced by cell death receptors is called the extrinsic pathway, and the one induced by mitochondria – the intrinsic pathway (Krammer et al., 2007). In both cases, the apoptosis signal initiates the activation of caspases, key enzymes of apoptosis, leading to cell destruction, but the molecular mechanisms of signal transmission are entirely different. The literature focuses on regulating the intrinsic pathway of apoptosis, in which there has been certain progress in finding compounds with pharmacological potential for HCC therapy. It should be noted that the pharmacological effect on the extrinsic apoptosis pathway in HCC remains poorly studied. However, pharmacological induction of this pathway may bring significant, fundamentally important progress for cancer therapy.

Apoptosis induction is controlled by a range of inhibitor proteins, including c-FLIP, which blocks the activation of caspase-8, members of the anti-apoptotic BCl-2 family that inhibit the release of cytochrome C from mitochondria, and XIAP proteins that block the activation of caspase-3, -7, and -9. In the extrinsic apoptosis pathway, DISC, comprising PC, FADD, procaspase-8, -10 proteins, and c-FLIP, serves as a central platform for procaspase-8 activation (Lavrik, Krammer, 2012). c-FLIP can function within the DISC complex both pro- and anti-apoptotically. It is suggested that the formation of procaspase-8/c-FLIP heterodimers mediates the pro-apoptotic function of c-FLIP. Previously, in joint research conducted by the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences and the University of Magdeburg, we developed the world's first chemical probe (small chemical compound) capable of specifically binding to c-FLIP in the caspase-8/c-FLIP heterodimeric complex (Hillert et al., 2020). This small molecule was obtained by computer design and possessed biological activity – the ability to increase caspase-8 activity (Hillert et al., 2020).

Hepatitis C virus (HCV) is extensively studied in the scientific literature as a significant risk factor for HCC (Axley et al., 2018). The role of HCV has been shown in the regulation

of apoptosis and aberrant gene methylation, closely associated with HCC (Zheng et al., 2019; Lee, Ou, 2021).

Gene networks are widely used to describe the molecular-genetic mechanisms of various processes. We previously developed the software and information system ANDSystem (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022), designed for the reconstruction and analysis of associative gene networks based on automatic knowledge extraction from scientific publications and factographic databases. Through the reconstruction of gene networks performed using ANDSystem, a number of studies have been conducted, such as the analysis of interactions of Hepatitis C virus proteins with the human proteome (Saik et al., 2016), the relationship of HCV with aberrant methylation in HCC (Antropova et al., 2022), interpretation of results of metabolome analysis of SARS-Cov-2 patients (Ivanisenko V.A. et al., 2022), tasks of prioritizing candidate genes associated with lymphedema, major depressive disorder (Yankina et al., 2018; Saik et al., 2019), search for new potential targets for drug action (Saik et al., 2018a, b), and others.

Based on the reconstruction and analysis of HCC gene networks and the extrinsic apoptosis pathway, as well as regulatory pathways linking HCV proteins with aberrantly methylated genes in HCC and key participants in the extrinsic apoptosis pathway, criteria were proposed for prioritizing potential pharmacological targets against HCC. Enrichment analysis of the first 100 target genes, ordered by prioritization results, showed significant content (p-value = 0.0002) in the list of FDA-approved pharmacological target genes, demonstrating the effectiveness of the proposed prioritization criteria. We suggest that the mechanism of action of drugs targeted at these targets is the modulation of the extrinsic apoptosis pathway, taking into account aberrant gene methylation, which could be utilized in creating a new class of drugs for HCC therapy. As promising potential pharmacological targets, ranked in the top thirty, the following candidate genes can be highlighted: *JUN*, *IL10*, *STAT3*, *MYC*, *TLR4*, and *KHDRBS1*.

Materials and methods

The ANDSystem software and information tool. Gene network reconstruction was performed using the ANDSystem software and information tool, automatically extracting knowledge from scientific publications and factual databases using artificial intelligence methods (Ivanisenko V.A. et al., 2019). ANDSystem includes a knowledge base containing over 40 million facts about molecular-genetic interactions, including physical intermolecular interactions, gene expression regulation, activity regulation, stability, and protein transport. Work on the reconstruction and analysis of gene networks in ANDSystem is performed using the ANDVisio program. The Pathway Wizard function implemented in ANDVisio was used to reconstruct regulatory pathways, which perform search queries to the knowledge base based on a given template. A schematic description of the templates used to reconstruct regulatory pathways is provided in Supplementary Materials 1–4[1].

Patient- and tissue-specific gene expression and DNA methylation data. Patient-specific and tissue-specific data on gene expression and DNA methylation were used to reconstruct gene networks. Tissue-specific gene expression data was used to filter gene networks using built-in ANDSystem methods. Information on tissue-specific gene expression was represented in ANDSystem. Information on differential gene expression was taken from the GEO database (Barrett et al., 2013; https://www.ncbi.nlm.nih.gov/geo/). Experiments were selected for which results of hepatocarcinoma tissue samples obtained from patients with this disease were available. The statistical significance values of differential gene expression and differential methylation in hepatocarcinoma tumor tissue samples compared to control samples were calculated in the GEO2R software package (Barrett et al., 2013; https://www.ncbi.nlm.nih.gov/geo/geo2r/). Calculation parameters were selected by default.

FDA-approved pharmacological targets. Data on FDA-approved pharmacological targets were extracted from the Human Protein Atlas resource (Uhlén et al., 2015; https://www.proteinatlas.org/).

Potential pharmacological target prioritization method. The criteria presented in Table 1 were used to prioritize candidate genes for pharmacological targets. The resulting gene weight was assessed as the sum of the weights of all criteria.

Results and discussion

To prioritize potential pharmacological targets, we applied 13 criteria considering various characteristics of the structural and functional organization of liver cancer gene networks and programmed cell death, including patient- and tissue-specific data on DNA methylation. Each criterion was assigned a quantitative weight indicator. The sum of the indicators for all 13 criteria was calculated as the resulting characteristic. To rank the genes by priority, they were arranged in a list from higher to lower values of the total indicator. Thus, genes with higher priority as candidates for pharmacological targets were at the top of the list (i. e., they had a lower rank).

When calculating the weight indicators of genes by prioritization criteria, the reconstruction of the gene networks of hepatocellular carcinoma (HCC) and the extrinsic apoptosis pathway was carried out as described below.

Reconstruction of the human hepatocellular carcinoma gene network

The automated search for genes associated with HCC, conducted using the new version of ANDSystem (Ivanisenko V.A. et al., 2019), identified more than 5,100 genes. Subsequently, ANDSystem built-in methods were used to filter genes by tissue specificity, retaining only the genes expressed in the liver – 4,905 genes. A list of 1,211 differentially expressed genes (DEGs) was then used based on RNA-seq analysis from the study by Huang et al. (2011). These data were obtained from the tissues of ten patients with HBV-associated HCC. Healthy tissues from the same patients were used as controls.

Following this step, the intersection of the gene network was reconstructed with ANDSystem, and the list of differentially expressed genes was carried out using ANDVisio built-in functions. As a result of the intersection, the gene network retained 584 genes found by ANDSystem methods to be associated with hepatocellular carcinoma based on data from published

[1] Supplementary Materials 1–7 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Demenkov_Engl_27_7.pdf

Table 1. Criteria developed for prioritizing candidate genes of pharmacological targets

No.	Criterion name	Value	Characteristic
1	Gene representation in the HCC gene network	score1 = 2	The gene or the protein it encodes is represented in the gene network
		score1 = 0	The gene or the protein it encodes is not represented in the gene network
2	Gene representation in the extrinsic apoptosis gene network	score2 = 2	The gene or the protein it encodes is represented in the gene network
		score2 = 0	The gene or the protein it encodes is not represented in the gene network
3	Aberrant methylation indicator	score3 = 3	The gene is hypomethylated in HCC (there is data on increased expression)
		score3 = –5	The gene is hypermethylated (there is data on decreased expression)
4	Centrality indicator of the gene in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD*) by genes from the HCC gene network (see Supplementary Material 1)	score4 = 1+ln(Q1)	The gene is represented in the regulatory gene network. Q1 – the number of connections of the gene with other nodes (degree centrality indicator)
		score4 = 0	The gene is not represented in the regulatory gene network
5	Centrality indicator of the protein in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD*) by genes from the HCC gene network (see Supplementary Material 1)	score5 = 1+ln(Q2)	The protein is represented in the regulatory gene network. Q2 – the number of connections of the protein with other nodes (degree centrality indicator)
		score5 = 0	The protein is not represented in the regulatory gene network
6	Centrality indicator of the gene in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD*) by HCV proteins (see Supplementary Material 2)	score6 = 2+ln(Q3)	The gene is represented in the regulatory gene network. Q3 – the number of connections of the gene with other nodes (degree centrality indicator)
		score6 = 0	The gene is not represented in the regulatory gene network
7	Centrality indicator of the protein in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD*) by HCV proteins (see Supplementary Material 2)	score7 = 2+ln(Q4)	The protein is represented in the regulatory gene network. Q4 – the number of connections of the protein with other nodes (degree centrality indicator)
		score7 = 0	The protein is not represented in the regulatory gene network
8	Centrality indicator of the gene in regulatory pathways (see Supplementary Material 3) describing the regulation of hypermethylated genes by HCV proteins	score8 = ln(Q5)	The gene is represented in the regulatory gene network. Q5 – the number of connections of the gene with other nodes (degree centrality indicator)
		score8 = 0	The gene is not represented in the regulatory gene network
9	Centrality indicator of the protein in regulatory pathways (see Supplementary Material 3) describing the regulation of hypermethylated genes by HCV proteins	score9 = ln(Q6)	The protein is represented in the regulatory gene network. Q6 – the number of connections of the protein with other nodes (degree centrality indicator)
		score9 = 0	The protein is not represented in the regulatory gene network
10	Centrality indicator of the gene in regulatory pathways (see Supplementary Material 3) describing the regulation of hypomethylated genes by HCV proteins	score10 = 1+ln(Q7)	The gene is represented in the regulatory gene network. Q7 – the number of connections of the gene with other nodes (degree centrality indicator)
		score10 = 0	The gene is not represented in the regulatory gene network
11	Centrality indicator of the protein in regulatory pathways (see Supplementary Material 3) describing the regulation of hypomethylated genes by HCV proteins	score11 = 1+ln(Q8)	The protein is represented in the regulatory gene network. Q8 – the number of connections of the protein with other nodes (degree centrality indicator)
		score11 = 0	The protein is not represented in the regulatory gene network
12	Centrality indicator of the gene in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD*) by aberrantly methylated genes (see Supplementary Material 4)	score12 = 2+ln(Q9)	The gene is represented in the regulatory gene network. Q9 – the number of connections of the gene with other nodes (degree centrality indicator)
		score12 = 0	The gene is not represented in the regulatory gene network
13	Centrality indicator of the protein in regulatory pathways describing the regulation of key genes of the extrinsic apoptosis pathway (*CFLAR*, *CASP8*, and *FADD)* by aberrantly methylated genes (see Supplementary Material 4)	score13 = 2+ln(Q10)	The protein is represented in the regulatory gene network. Q10 – the number of connections of the protein with other nodes (degree centrality indicator)
		score13 = 0	The protein is not represented in the regulatory gene network

works and databases, which were also present in the list of differentially expressed genes of human hepatocellular carcinoma obtained from RNA-seq data in (Huang et al., 2011). A search was then conducted for proteins expressed from these genes and metabolites associated with these proteins through direct interactions (a 'catalyst' type association), and a network of interactions between all objects in the gene network (genes, proteins, and metabolites) was reconstructed. The gene network contained 584 genes, 580 proteins, 1,061 metabolites, and over 16,000 interactions at this stage.

The gene network was expanded in the second stage with patient- and tissue-specific DNA methylation data (Supplementary Material 5). This included 67 genes with differentially altered methylation (hyper- or hypomethylated genes) in patient tumors compared to control samples. After adding aberrantly methylated genes and their protein products and expanding the gene network with metabolites interacting with them, the final gene network contained 627 genes, 624 proteins, 1,105 metabolites, and 17,387 interactions.

Reconstruction of the extrinsic apoptosis pathway gene network

The gene network of the extrinsic apoptosis pathway was reconstructed considering GeneOntology and ANDSystem data (Supplementary Material 6). Initially, a list of genes involved in the extrinsic apoptotic signaling pathway was formed using a query to the GeneOntology database. The following keywords were used for the query: GO term "extrinsic apoptotic signaling pathway", organism "human". Based on this query, a list of 259 genes was obtained. This list was then uploaded into the ANDVisio program to construct a gene network. Using ANDSystem, the gene network was expanded with proteins expressed from the entered genes, as well as with metabolites associated with these genes. As a result, the gene network of the extrinsic apoptosis pathway contained 259 genes, 260 proteins, and 513 metabolites.

Gene prioritization results

A total of 1,345 genes were analyzed, including participants in the HCC and extrinsic apoptosis pathway gene networks and regulatory pathways. The results of applying prioritization criteria for the top 30 priority genes are presented in Table 2. Out of 1,345 genes, 137 were targets of FDA-approved drugs. The top 100 priority list included 19 genes targeted by FDA-approved drugs. Detailed information on the results of prioritization for the 100 highest priority genes, containing quantitative values for each of the criteria, is provided in Supplementary Material 7. Of these 19 target genes, 17 are characterized as cancer-related genes. According to the hypergeometric distribution, the probability of an event in which 17 or more out of 19 selected genes are associated with cancer is $p = 0.0002$. This analysis signifies that the top 100 priority genes in the table of potential targets are statistically significantly associated with cancer (significance level $p = 0.0002$).

The calculation of prioritization criteria indicators, based on the reconstruction of regulatory pathways (criteria 4–13), was conducted automatically using ANDSystem with the templates provided in Supplementary Materials 1–4. The reconstruction and analysis of regulatory pathways of hypermethylated genes by Hepatitis C viral proteins, the results of which were used in prioritization criteria 8–11, have been previously described by us (Antropova et al., 2022).

The *JUN* gene occupies the top rank in the table (see Table 2). It belongs to the group of drug target genes approved by the FDA and is also associated with cancer (cancer-related genes). Numerous literature reports discuss its role in various types of cancer. For instance, it has been shown that JUN affects the development of colon cancer (Nateri et al., 2005) and that activated JUN is predominantly expressed at the invasive front of breast cancer and is associated with proliferation and angiogenesis (Vleugel et al., 2006).

According to our results, this gene could regulate the extrinsic apoptosis pathway. The regulatory network we reconstructed, which describes the molecular pathways through which JUN could regulate the extrinsic apoptosis pathway markers CFLAR, CASP8, and FADD, is presented in Figure 1. The regulatory network is based on various conclusions from experimental studies. For example, it has been shown that *FASLG* expression depends on JUN – irradiation increased FASLG expression in GCK cells via the activation of the JNK/c-Jun signaling pathway (Dong et al., 2016). The *FASLG* gene encodes the TNFL6 protein, a cytokine that binds to the TNFRSF6/FAS receptor, transmitting an apoptosis signal to cells. In another study (Liu Z. et al., 2019), deletion of *FASLG* inhibited the expression of *CASP8*, demonstrating another possible way for JUN to influence apoptosis (via CASP8).

It should be noted that pharmacological targets approved by the FDA, which are not associated with cancer but may be related to apoptosis, also present a particular interest. Specifically, in our table, *TLR4* (ranked 14th) stands out among such genes. According to the FDA, the *TLR4* gene is associated with "age-related macular degeneration" disease. Disruption of apoptosis is a key pathological factor in this disease (Yi et al., 2012).

The regulatory network describing the molecular pathways through which TLR4 can regulate CFLAR, CASP8, and FADD is presented in Figure 2. For instance, one can observe the regulatory influence from TLR4 to *TNFAIP3*. It is reconstructed based on a published study, showing that TLR4 activates a signaling pathway leading to the activation of NF-κB transcription factor. NF-κB, in turn, induces the expression of *TNFAIP3*, as demonstrated in endothelial cells (Soni et al., 2018). TNFAIP3 increases the level of cleaved caspase-8, as confirmed by knockdown, while overexpression of *TNFAIP3* has the opposite effect (Liu K. et al., 2018). Similarly, TLR4 could enhance the expression of *Beclin-1* through NF-κB (Copetti et al., 2009), which induces caspase-8 cleavage, leading to autophagy and apoptosis (Song et al., 2014).

The *IL10* gene occupies the second rank in the table. It belongs to the group of genes not included in the list of FDA-approved pharmacological targets. However, their mechanisms of influence on the development of HCC are widely discussed in the literature. In 2020, a study (Qian et al., 2020) suggested that combining IL10 and PD-L1 inhibitors may form the basis for effective treatment. The regulatory network, describing the molecular pathways through which IL10 can regulate CFLAR, CASP8, and FADD, is presented in Figure 3.

Another group consists of genes for which the FDA does not indicate approved agents, yet the mechanism of action of

Table 2. Top 30 genes ranked by priority level

Rank	Gene	Full gene name	Presence of FDA-approved* agents	Total weight
1	*JUN*	Proto-oncogene c-Jun	CR**	37.4
2	*IL10*	Interleukin-10	–	30.9
3	*STAT3*	Signal transducer and activator of transcription 3	–	30.1
4	*CASP8*	Caspase-8	–	29.4
5	*TP53*	Cellular tumor antigen p53	–	28.7
6	*CFLAR*	CASP8 and FADD-like apoptosis regulator	–	28.3
7	*MYC*	Myc proto-oncogene protein	–	23.7
8	*NFKB1*	Nuclear factor NF-kappa-B p105 subunit	CR	23.2
9	*FADD*	FAS-associated death domain protein	–	23.0
10	*IL33*	Interleukin-33	–	23.0
11	*ELAVL1*	ELAV-like protein 1	–	22.9
12	*FASLG*	Tumor necrosis factor ligand superfamily member 6	–	22.8
13	*TERT*	Telomerase reverse transcriptase	–	22.5
14	*TLR4*	Toll-like receptor 4	AR***	22.4
15	*BECN1*	Beclin-1	–	22.3
16	*CLDN1*	Claudin-1	–	22.3
17	*PARP1*	Poly [ADP-ribose] polymerase 1	CR	22.3
18	*TNFRSF1A*	Tumor necrosis factor receptor superfamily member 1A	CR	21.8
19	*CDKN1A*	Cyclin-dependent kinase inhibitor 1	–	21.6
20	*SP1*	Transcription factor Sp1	–	21.1
21	*KHDRBS1*	KH domain-containing, RNA-binding, signal transduction-associated protein 1	–	20.6
22	*MCL1*	Induced myeloid leukemia cell differentiation protein	–	20.6
23	*CLDN7*	Claudin-7	–	20.3
24	*CTSD*	Cathepsin D	–	20.0
25	*FASN*	Fatty acid synthase	CR	19.1
26	*MYCN*	N-myc proto-oncogene protein	–	18.7
27	*DDIT3*	DNA damage-inducible transcript 3 protein	–	18.4
28	*TNFAIP3*	Tumor necrosis factor alpha-induced protein 3	–	18.1
29	*STAT1*	Signal transducer and activator of transcription 1	–	17.6
30	*NLRP3*	NACHT, LRR and PYD domains-containing protein 3	–	17.6

* FDA – Food and Drug Administration, the agency of the US Department of Health and Human Services responsible for the sanitary supervision of food products and medicines; ** CR – cancer-related genes; *** AR – genes related to the "age-related macular degeneration" disease.

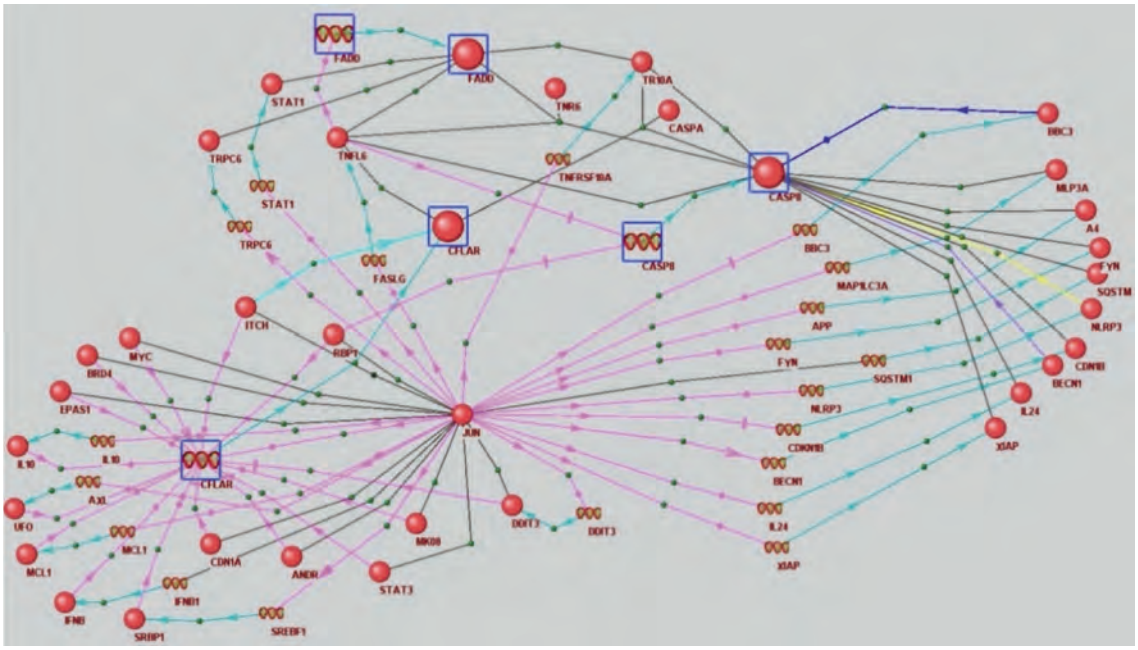


Fig. 1. Interaction network reconstructed using ANDSystem, through which JUN can regulate key apoptosis proteins – CFLAR, CASP8, and FADD.

Spheres represent proteins, and spirals symbolize genes. Black lines indicate physical interaction, turquoise arrows denote expression, pink arrows signify regulation of expression, blue arrows represent transport regulation, and yellow arrows indicate activity regulation.

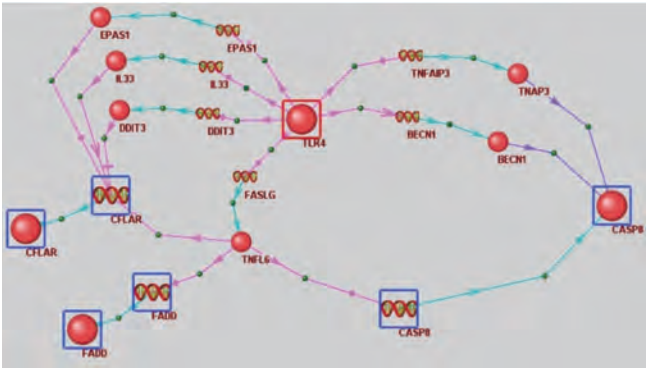


Fig. 2. Interaction network reconstructed using ANDSystem, through which TLR4 can regulate key apoptosis proteins – CFLAR, CASP8, and FADD.

Spheres represent proteins, and spirals symbolize genes. Turquoise arrows indicate expression, purple arrows represent regulation, and pink arrows denote expression regulation.

some widely used drugs affects these genes or the proteins they encode. This group includes the *STAT3* and *MYC* genes, occupying the rank table's third and seventh positions. A substantial number of publications indicate that STAT3 plays a crucial role in the initiation, progression, immune suppression, and metastasis of HCC. Specific drugs affect the functioning of STAT3. For instance, F.M. Gu et al. demonstrated that the inhibition of HCC growth and metastasis by the targeted anticancer drug "sorafenib" is mediated by blocking STAT3 (Gu et al., 2011). It is also known that sorafenib induces apoptosis (Xie et al., 2012). L. Wu et al., studying the mechanism of action of quercetin (a natural flavonoid included in some dietary supplements and drugs), showed that it inhibits the progression of HCC, affecting apoptosis, migration, invasion, autophagy, via the JAK2/STAT3 signaling pathway (at least partially) (Wu et al., 2019). The action mechanism of another anticancer drug – trametinib, used for melanoma treatment, is based on inhibiting the MEK protein, part of the signaling cascade. MEK inhibition reduces the MYC protein level, which promotes cell survival, and increases the pro-apoptotic protein BIM level, suppressing HCC growth (Zhou et al., 2019).

The direct markers of the extrinsic apoptosis pathway, *CASP8*, and *CFLAR*, are ranked 4th and 6th in the rank table. The *TP53* gene, the importance of which for apoptosis is well known, is positioned between them at the fifth position. Thus, it can be concluded that among the potential pharmacological targets we found, the top results of prioritization (see Table 2) include genes that are indeed drug targets – either FDA-approved or drugs aimed at other targets but affecting these genes and the proteins they encode in their action mechanisms, as well as genes that are only currently being discussed as promising targets.

Of particular interest as pharmacological targets may be genes that have been poorly studied to date in relation to HCC development mechanisms. Such genes could be fundamentally new pharmacological targets. Specifically, among such genes that made it to the top 100 highest priority list is *KHDRBS1*, which occupies the 21st position in the rank table (see Table 2). The regulatory network describing the molecular pathways through which KHDRBS1 can regulate CFLAR, CASP8, and FADD is presented in Figure 4.

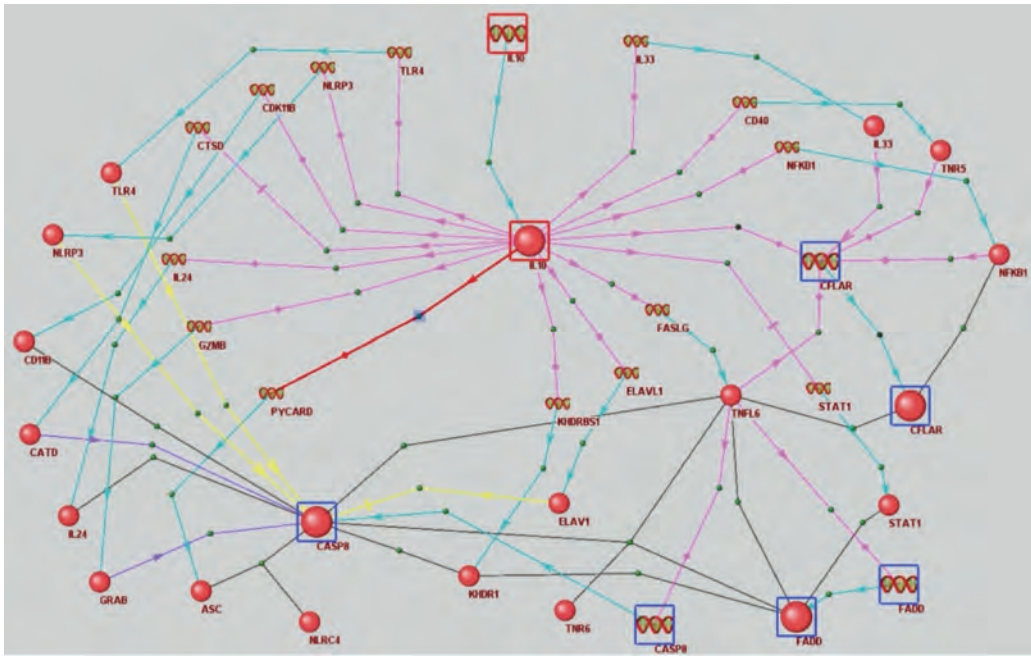


Fig. 3. Interaction network reconstructed using ANDSystem, through which IL10 can influence CFLAR, CASP8, and FADD.

Spheres represent proteins, and spirals symbolize genes. Black lines indicate physical interaction, turquoise arrows denote expression, purple arrows represent regulation, pink arrows signify regulation of expression, and yellow arrows indicate activity regulation.

Conclusion

A computer reconstruction of gene networks for hepatocellular carcinoma and programmed cell death (extrinsic apoptosis pathway) was conducted, taking into account patient- and tissue-specific DNA methylation data, using the ANDSystem software and information system. Based on the 13 developed criteria, considering the specifics of the reconstructed gene networks' structural and functional organization, potential pharmacological targets were prioritized. Six candidate genes (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4*, and *KHDRBS1*), occupying high positions in the ranked list according to prioritization results, may be of greatest interest as potential pharmacological targets.

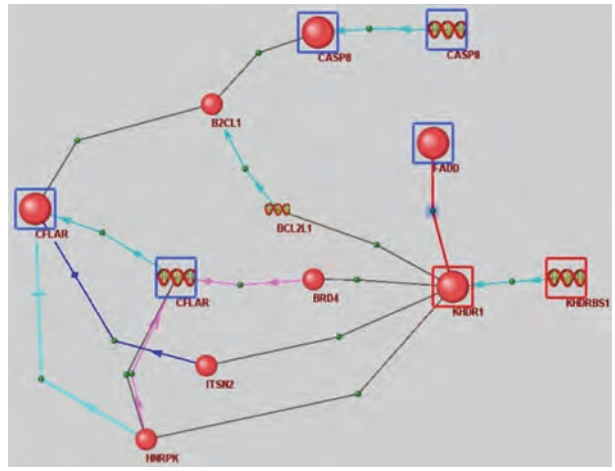


Fig. 4. Interaction network reconstructed using ANDSystem, through which KHDRBS1 can regulate key apoptosis proteins – CFLAR, CASP8, and FADD.

Spheres represent proteins, and spirals symbolize genes. Black lines indicate physical interaction, turquoise arrows denote expression, and pink arrows signify expression regulation.

References

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Lavrik I.N., Ivanisenko V.A. Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. DOI 10.18699/VJGB-22-89

Axley P., Ahmed Z., Ravi S., Singal A.K. Hepatitis C virus and hepatocellular carcinoma: a narrative review. *J. Clin. Transl. Hepatol*. 2018;6(1):79-84. DOI 10.14218/JCTH.2017.00067

Balogh J., Victor D., Asham E.H., Burroughs S.G., Boktour M., Saharia A., Li X., Ghobrial R.M., Monsour H.P., Jr. Hepatocellular carcinoma: a review. *J. Hepatocell. Carcinoma*. 2016;3:41-53. DOI 10.2147/JHC.S61146

Barrett T., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C.L., Serova N., Davis S., Soboleva A. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41(D1):D991-D995. DOI 10.1093/nar/gks1193

Copetti T., Bertoli C., Dalla E., Demarchi F., Schneider C. p65/RelA modulates BECN1 transcription and autophagy. *Mol. Cell. Biol*. 2009;29(10):2594-2608. DOI 10.1128/MCB.01396-08

Dong Y., Shen X., He M., Wu Z., Zheng Q., Wang Y., Chen Y., Wu S., Cui J., Zeng Z. Activation of the JNK-c-Jun pathway in response to irradiation facilitates Fas ligand secretion in hepatoma cells and in-

creases hepatocyte injury. *J. Exp. Clin. Cancer Res*. 2016;35(1):114. DOI 10.1186/s13046-016-0394-z

Forner A., Llovet J.M., Bruix J. Hepatocellular carcinoma. *Lancet*. 2012;379(9822):1245-1255. DOI 10.1016/S0140-6736(11)61347-0

Gu F.M., Li Q.L., Gao Q., Jiang J.H., Huang X.Y., Pan J.F., Fan J., Zhou J. Sorafenib inhibits growth and metastasis of hepatocellular carcinoma by blocking STAT3. *World J. Gastroenterol*. 2011; 17(34):3922-3932. DOI 10.3748/wjg.v17.i34.3922

Hillert L.K., Ivanisenko N.V., Busse D., Espe J., König C., Peltek S.E., Kolchanov N.A., Ivanisenko V.A., Lavrik I.N. Dissecting DISC regulation via pharmacological targeting of caspase-8/c-FLIP$_L$ heterodimer. *Cell Death Differ*. 2020;27(7):2117-2130. DOI 10.1038/s41418-020-0489-0

Huang Q., Lin B., Liu H., Ma X., Mo F., Yu W., Li L., Li H., Tian T., Wu D., Shen F., Xing J., Chen Z.N. RNA-seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS One*. 2011;6(10):e26168. DOI 10.1371/journal.pone.0026168

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(Suppl.11):228. DOI 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int. J. Mol. Sci.* 2022;23(23):14934. DOI 10.3390/ijms232314934

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol*. 2015;9(Suppl.2):S2. DOI 10.1186/ 1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. DOI 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci. Rep*. 2022;12(1):19977. DOI 10.1038/s41598-022-24170-0

Krammer P.H., Kamiński M., Kiessling M., Gülow K. No life without death. *Adv. Cancer Res*. 2007;97:111-138. DOI 10.1016/S0065-230X(06)97005-5

Lavrik I.N., Krammer P.H. Regulation of CD95/Fas signaling at the DISC. *Cell Death Differ*. 2012;19(1):36-41. DOI 10.1038/cdd.2011.155

Lee J., Ou J.J. Hepatitis C virus and intracellular antiviral response. *Curr. Opin. Virol*. 2022;52:244-249. DOI 10.1016/j.coviro.2021.12.010

Liu K., Yao H., Wen Y., Zhao H., Zhou N., Lei S., Xiong L. Functional role of a long non-coding RNA LIFR-AS1/miR-29a/TNFAIP3 axis in colorectal cancer resistance to pohotodynamic therapy. *Biochim. Biophys. Acta Mol. Basis Dis*. 2018;1864(9B):2871-2880. DOI 10.1016/j.bbadis.2018.05.020

Liu Z., Fitzgerald M., Meisinger T., Batra R., Suh M., Greene H., Penrice A.J., Sun L., Baxter B.T., Xiong W. CD95-ligand contributes to abdominal aortic aneurysm progression by modulating inflammation. *Cardiovasc. Res*. 2019;115(4):807-818. DOI 10.1093/cvr/cvy264

Llovet J.M., Montal R., Sia D., Finn R.S. Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol*. 2018;15(10):599-616. DOI 10.1038/s41571-018-0073-4

Nateri A.S., Spencer-Dene B., Behrens A. Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature*. 2005;437(7056):281-285. DOI 10.1038/nature03914

Qian Q., Wu C., Chen J., Wang W. Relationship between IL10 and PD-L1 in liver hepatocellular carcinoma tissue and cell lines. *Biomed. Res. Int*. 2020;2020:8910183. DOI 10.1155/2020/8910183

Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res*. 2016;218:40-48. DOI 10.1016/j.virusres.2015.12.003

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Dosenko V.E., Zolotareva O.I., Choynzonov E.L., Hofestaedt R., Ivanisenko V.A. Search for new candidate genes involved in the comorbidity of asthma and hypertension based on automatic analysis of scientific literature. *J. Integr. Bioinform*. 2018a;15(4):20180054. DOI 10.1515/jib-2018-0054

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofestaedt R., Lavrik I.N., Rogaev E.I. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018b;11(1):61-76. DOI 10.1186/s12920-018-0331-4

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med. Genomics*. 2019;12(Suppl.2):117-131. DOI 10.1186/s12920-019-0492-9

Song X., Kim S.Y., Zhang L., Tang D., Bartlett D.L., Kwon Y.T., Lee Y.J. Role of AMP-activated protein kinase in cross-talk between apoptosis and autophagy in human colon cancer. *Cell Death Dis*. 2014;5(10):e1504. DOI 10.1038/cddis.2014.463

Soni D., Wang D.M., Regmi S.C., Mittal M., Vogel S.M., Schlüter D., Tiruppathi C. Deubiquitinase function of A20 maintains and repairs endothelial barrier after lung vascular injury. *Cell Death Discov*. 2018;4:60. DOI 10.1038/s41420-018-0056-3

Thorgeirsson S.S., Grisham J.W. Molecular pathogenesis of human hepatocellular carcinoma. *Nat. Genet*. 2002;31(4):339-346. DOI 10.1038/ng0802-339

Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., Kampf C., Sjöstedt E., Asplund A., Olsson I., Edlund K., Lundberg E., Navani S., Szigyarto C.A., Odeberg J., Djureinovic D., Takanen J.O., Hober S., Alm T., Edqvist P.H., Berling H., Tegel H., Mulder J., Rockberg J., Nilsson P., Schwenk J.M., Hamsten M., von Feilitzen K., Forsberg M., Persson L., Johansson F., Zwahlen M., von Heijne G., Nielsen J., Pontén F. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. DOI 10.1126/science.1260419

Vleugel M.M., Greijer A.E., Bos R., van der Wall E., van Diest P.J. c-Jun activation is associated with proliferation and angiogenesis in invasive breast cancer. *Hum. Pathol*. 2006;37(6):668-674. DOI 10.1016/j.humpath.2006.01.022

Wu L., Li J., Liu T., Li S., Feng J., Yu Q., Zhang J., Chen J., Zhou Y., Ji J., Chen K., Mao Y., Wang F., Dai W., Fan X., Wu J., Guo C. Quercetin shows anti-tumor effect in hepatocellular carcinoma LM3 cells by abrogating JAK2/STAT3 signaling pathway. *Cancer Med*. 2019;8(10):4806-4820. DOI 10.1002/cam4.2388

Xie B., Wang D.H., Spechler S.J. Sorafenib for treatment of hepatocellular carcinoma: a systematic review. *Dig. Dis. Sci*. 2012;57(5):1122-1129. DOI 10.1007/s10620-012-2136-1

Yankina M.A., Saik O.V., Ivanisenko V.A., Demenkov P.S., Khusnutdinova E.K. Evaluation of prioritization methods of extrinsic apoptotic signaling pathway genes for retrieval of the new candidates associated with major depressive disorder. *Russ. J. Genet*. 2018; 54(11):1366-1374. DOI 10.1134/S1022795418110170

Yi H., Patel A.K., Sodhi C.P., Hackam D.J., Hackam A.S. Novel role for the innate immune receptor Toll-like receptor 4 (TLR4) in the regulation of the Wnt signaling pathway and photoreceptor apoptosis. *PLoS One*. 2012;7(5):e36560. DOI 10.1371/journal.pone.0036560

Zheng Y., Hlady R.A., Joyce B.T., Robertson K.D., He C., Nannini D.R., Kibbe W.A., Achenbach C.J., Murphy R.L., Roberts L.R., Hou L. DNA methylation of individual repetitive elements in hepatitis C virus infection-induced hepatocellular carcinoma. *Clin. Epigenetics*. 2019;11(1):145. DOI 10.1186/s13148-019-0733-y

Zhou X., Zhu A., Gu X., Xie G. Inhibition of MEK suppresses hepatocellular carcinoma growth through independent MYC and BIM regulation. *Cell. Oncol.* (*Dordr*.). 2019;42(3):369-380. DOI 10.1007/s13402-019-00432-4

ORCID ID

P.S. Demenkov orcid.org/0000-0001-9433-8341
E.A. Antropova orcid.org/0000-0003-2158-3252
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
A.S. Venzel orcid.org/0000-0002-7419-5168
V.A. Ivanisenko orcid.org/0000-0002-1859-4631

Acknowledgements. This research was conducted with the financial support of project No. 075-15-2021-944 by the Ministry of Science and Higher Education of the Russian Federation within the framework of the ERA-NET "Target Identification and Drug Development in Liver Cancer (TAIGA)".

Conflict of interest. The authors declare no conflict of interest.

Received July 26, 2023. Revised August 25, 2023. Accepted August 28, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):794-806
DOI 10.18699/VJGB-23-92

Original Russian text https://vavilovj-icg.ru/

RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research

I.V. Chadaeva[1], S.V. Filonov[1, 2], K.A. Zolotareva[1], B.M. Khandaev[1], N.I. Ershov[1], N.L. Podkolodnyy[1, 3], R.V. Kozhemyakina[1], D.A. Rasskazov[1], A. G. Bogomolov[1], E.Yu. Kondratyuk[1, 4], N.V. Klimova[1], S.G. Shikhevich[1], M.A. Ryazanova[1], L.A. Fedoseeva[1], O.E. Redina[1], O.S. Kozhevnikova[1], N.A. Stefanova[1], N.G. Kolosova[1], A.L. Markel[1, 2], M.P. Ponomarenko[1] ✉, D.Yu. Oshchepkov[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, Russia
✉ pon@bionet.nsc.ru

Abstract. The animal models used in biomedical research cover virtually every human disease. RatDEGdb, a knowledge base of the differentially expressed genes (DEGs) of the rat as a model object in biomedical research is a collection of published data on gene expression in rat strains simulating arterial hypertension, age-related diseases, psychopathological conditions and other human afflictions. The current release contains information on 25,101 DEGs representing 14,320 unique rat genes that change transcription levels in 21 tissues of 10 genetic rat strains used as models of 11 human diseases based on 45 original scientific papers. RatDEGdb is novel in that, unlike any other biomedical database, it offers the manually curated annotations of DEGs in model rats with the use of independent clinical data on equal changes in the expression of homologous genes revealed in people with pathologies. The rat DEGs put in RatDEGdb were annotated with equal changes in the expression of their human homologs in affected people. In its current release, RatDEGdb contains 94,873 such annotations for 321 human genes in 836 diseases based on 959 original scientific papers found in the current PubMed. RatDEGdb may be interesting first of all to human geneticists, molecular biologists, clinical physicians, genetic advisors as well as experts in biopharmaceutics, bioinformatics and personalized genomics. RatDEGdb is publicly available at https://www.sysbio.ru/RatDEGdb.
Key words: knowledge base; DEG; *Rattus norvegicus*; animal models of human diseases; neurodegeneration; Alzheimer's disease; hypertension; premature aging; psychopathological states; catatonic syndrome; epilepsy; aggression; RNA-seq; PCR; microarrays.

For citation: Chadaeva I.V., Filonov S.V., Zolotareva K.A., Khandaev B.M., Ershov N.I., Podkolodnyy N.L., Kozhemyakina R.V., Rasskazov D.A., Bogomolov A.G., Kondratyuk E.Yu., Klimova N.V., Shikhevich S.G., Ryazanova M.A., Fedoseeva L.A., Redina O.E., Kozhevnikova O.S., Stefanova N.A., Kolosova N.G., Markel A.L., Ponomarenko M.P., Oshchepkov D.Yu. RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):794-806. DOI 10.18699/VJGB-23-92

База знаний RatDEGdb по дифференциально экспрессирующимся генам крысы как модельного объекта биомедицинских исследований

И.В. Чадаева[1], С.В. Филонов[1, 2], К.А. Золотарева[1], Б.М. Хандаев[1, 2], Н.И. Ершов[1], Н.Л. Подколодный[1, 3], Р.В. Кожемякина[1], Д.А. Рассказов[1], А.Г. Богомолов[1], Е.Ю. Кондратюк[1, 4], Н.В. Климова[1], С.Г. Шихевич[1], М.А. Рязанова[1], Л.А. Федосеева[1], О.Е. Редина[1], О.С. Кожевникова[1], Н.А. Стефанова[1], Н.Г. Колосова[1], А.Л. Маркель[1, 2], М.П. Пономаренко[1] ✉, Д.Ю. Ощепков[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Сибирский федеральный научный центр агробиотехнологий Российской академии наук, р.п. Краснообск, Новосибирская область, Россия
✉ pon@bionet.nsc.ru

Аннотация. Животные модели, используемые в биомедицинских исследованиях, в настоящее время охватывают практически весь известный спектр заболеваний человека. База знаний RatDEGdb по дифференциально экспрессирующимся генам (ДЭГ) крысы как модельного объекта в биомедицинских исследованиях представляет собой коллекцию опубликованных данных по экспрессии генов у крыс разных линий, предназначенных для изучения артериальной гипертонии, болезней пожилого возраста, психопатологических состояний и других заболеваний

человека. Текущий выпуск RatDEGdb содержит 25101 ДЭГ, представляющих 14320 уникальных генов крысы, которые изменяют уровень транскрипции в 21 ткани 10 генетических линий крысы в качестве моделей 11 заболеваний человека согласно 45 оригинальным научным статьям. Новшество RatDEGdb по сравнению с другими биомедицинскими базами данных заключается в курируемой аннотации отклонений ДЭГ крысы как модельного объекта с использованием независимых клинических данных об однонаправленных изменениях экспрессии гомологичных генов, выявленных у людей при различных патологиях. Собранные ДЭГ крыс были аннотированы однонаправленными изменениями экспрессии гомологичных им генов человека у больных людей относительно здоровых. К настоящему времени выпуск RatDEGdb содержит 94873 такие аннотации для 321 гена человека при 836 заболеваниях согласно 959 оригинальным научным статьям, найденным в текущем выпуске базы данных PubMed. Представленная база знаний может быть интересна в первую очередь специалистам по генетике человека, молекулярным биологам, клиницистам и генетическим консультантам, а также специалистам в области биофармацевтики, биоинформатики и персонализированной геномики. RatDEGdb является общедоступной (https://www.sysbio.ru/RatDEGdb).
Ключевые слова: база знаний; ДЭГ; крысы *Rattus norvegicus*; животные модели болезней человека; нейродегенерация; болезнь Альцгеймера; гипертоническая болезнь; преждевременное старение; психопатологические состояния; кататонический синдром; эпилепсия; агрессивность; RNA-seq; ПЦР; микрочипы.

Introduction

The animal models required for understanding the physiological, genetic and epigenetic mechanisms regulating evolutionarily fixed phenotypic traits of an organism are supposed to perfectly mimic the symptoms of the pathology being studied and to conform to strict criteria (Gryksa et al., 2023). The most popular animal models are rats and mice, with dozens of thousands of laboratory strains in use (Gayday E.A., Gayday D.S., 2019).

The first inbred rat strain was developed in 1906 in the Wistar Institute (Philadelphia, USA), about the time that mice came to the laboratory settings. Nevertheless, the mouse has become the model of choice for research into mammalian genetics, and the rat, into physiology and biomedicine. Laboratory rats have certain advantages over mice: rats are larger and therefore submit more tissue for analyses. Large organs make surgical procedures more manageable and rather small anatomical structures easier to dissect.

A low maintenance and cheap species, the rat (*Rattus norvegicus*) has become a convenient object in numerous biomedical research studies (Carter et al., 2020; Modlinska, Pisula, 2020). Rats are recommended for use as model animals in studying aging, hypertension, catalepsy etc. (Carter et al., 2020; Martín-Carro et al., 2023).

There are generally acknowledged differences between wild and laboratory rats. For example, laboratory rats are noted for smaller adrenals and preputial glands, earlier puberty, lack of seasonality of reproduction and higher fertility than have their wild conspecifics. In addition, the rat and human genomes share a 90 % identity (Gibbs et al., 2004). Thus, the genetic strains of laboratory rats simulating human pathologies have been developed: for example, the Zucker strain for human obesity, hypertension, type II diabetes and heart disease (Schmidt, 2002); the reelin-deficient shaking rat Kawasaki for schizophrenia and autism (Aikawa et al., 1988); and the Brattleboro strain for hypothalamic diabetes insipidus (Ideno et al., 2003). To date, there are about 1,000 inbred strains of laboratory rats developed by genetic breeding that have "fixed" alleles for natural diseases (Greenhouse et al., 1990), such as mental disorders (Taylor et al., 2002), depression (Bay et al., 2020) and chronic renal failure (Zhang H.F. et al., 2019). The Wistar and Sprague-Dawley strains are the most commonly used laboratory rats (Sengupta, 2013). At present, the search of PubMed (Lu, 2011) with "rats biomedical model" as a search string returns the annotations of 19,555 original scientific papers, which lends support to the relevance of the subject.

To contribute to the effort, several rat strains simulating human diseases have been developed in the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Thus, the ISIAH rats are characterized by an increased arterial blood pressure and used for studying the causes and treatments of hypertension in humans (Markel, 1992; Markel et al., 1999; Fedoseeva et al., 2016a, 2019; Klimov et al., 2016; Ryazanova et al., 2016), the OXYS rats represent a unique selection-based model of premature ageing and associated diseases (Kozhevnikova et al., 2013; Kolosova et al., 2014; Perepechaeva et al., 2014; Stefanova et al., 2018, 2019; Stefanova, Kolosova, 2023), rats with pendulum-like movements (the PM strain) with stereotypies and audiogenic epilepsy, and rats with genetic catatonia (the GC rats), a syndrome observed in patients with mental disorders, including schizophrenia (Barykina et al., 1983; Kolpakov et al., 2004; Ryazanova et al., 2017, 2023).

Changes in the expression of the genes associated with a disease of interest have been studied in the model rats by semi-quantitative real-time PCR of separate key genes or by profiling transcriptomes by next-generation sequencing or by use of microarrays. This effort has created a large body of data on the differentially expressed genes (DEGs) significantly associated with diseases, and it has become possible to collect, perform comparative analyses on and systematize the results obtained from these or similar experiments with the use of bioinformatics technologies. This has enabled the development of specialized databases and knowledge bases.

The aim of this work was to create a knowledge base containing information on DEGs of various rat strains developed, first of all, in the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences as well as those developed in a range of Russia's and other scientific organizations. This knowledge base is freely available at https://www.sysbio.ru/RatDEGdb.

Materials and methods

Experimental animals. We performed *in vivo* experiments on 12 adult male gray rat (*Rattus norvegicus*) from two outbred strains resulting from genetic breeding for more than 90 generations in two directions (Belyaev, Borodin, 1982): one for increased aggressive behavior towards humans (the aggressive strain) and one for decreased (the tame strain). The animals were kept in standard conditions at the Conventional Animal Facility of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia) as groups by four in 50×33×20 cm cages at an adjustable light/dark cycle (12 light:12 dark) and had free access to water and complete feed.

The test subjects were two-month-old individuals, each weighing 250–270 g, from unrelated litters. Within the first 4 hours of the light phase of the diurnal light-dark cycle, each animal's level of tameness/aggression was measured in the "glove" test as the reaction to a gloved hand and was scored from "–4" (most aggressive) to "+4" (most friendly), according to Plyusnina and Oskina (1997). Upon the completion of this test, the animals were put back to their home cages and kept in standard conditions for one week, to reduce possible effects that the "glove" test might have on gene expression, at which point the animals were euthanized and hypothalamus specimens were prepared according to the brain atlas of Paxinos and Watson (2013). Samples were placed in liquid nitrogen for transportation and further storage at –70 °C until use. The protocol of experiments was approved by the Commission on Bioethics at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (resolution No. 97 as of October 28, 2021).

Measurement of the hypothalamic mRNA levels of the *Asmtl* gene in tame and aggressive male gray rats by semi-quantitative PCR. To measure mRNA levels by semi-quantitative real-time polymerase chain reaction, hypothalamic RNA was isolated from six aggressive rats ($n = 6$) and six tame rats ($n = 6$), each specimen weighing ~100 mg. Total RNA was isolated using TRIzol™ (Invitrogen, #15596018) and purified using magnetic beads in the Agencourt RNAClean XP Kit (Beckman, #A63987). Purified RNA was quantified using a Qubit™ 2.0 fluorimeter (Invitrogen/Life Technologies) and a Qubit™ RNA High-Sensitivity Assay Kit (Invitrogen #In=Q32852). Next, we synthesized cDNA using the Reverse Transcription Kit (Syntol, #OT-1).

The oligonucleotide primers for each gene in question were designed using the web service PrimerBLAST (Ye et al., 2012) (Table 1). Real-time PCR was carried out using the EVA Green I kit in three technical replicates in a LightCycler® 96 operated in the automatic mode, according to the manufacturer's instruction (Roche, Switzerland). The efficiency of the polymerase chain reaction was determined by serial cDNA dilutions (standards).

The human gene *ASMT* encodes acetylserotonin O-methyltransferase, a key enzyme in the synthesis of melatonin, one of the hormones that regulate the molecular and genetic processes in the entire organism, including circadian rhythms as well as cancer protective (Lv et al., 2019), anti-inflammatory, and immunomediatory mechanisms (Li G. et al., 2021). That is why the mRNA level of its rat homolog, *Asmtl*, in the hypothalamus of adult tame and aggressive male rats used as model animals in the biomedical studies of increased aggression was heuristically chosen as the quantity to be found by semi-quantitative real-time PCR (real-time PCR) in its first run. As was recommended by Bustin and the co-workers (2009), the *Asmtl* mRNA values were normalized to the mRNA levels of two comparison genes, *Ppia* (Gholami et al., 2017) and *Rpl30* (Penning et al., 2007). The relevance of *Ppia* and *Rpl30* as the comparison genes in the experimental identification of DEGs in the hypothalamus of these aggressive and tame rat strains by real-time PCR was demonstrated in one of our previous works (Chadaeva et al., 2021).

RatDEGdb: the knowledge base. The observed lower hypothalamic levels of the *Asmtl* gene in the adult aggressive and tame male rats were checked against clinical data suggesting that lower levels of the protein encoded by its human homologs *ASMT* and *ASMTL* were in patients with various diseases than in otherwise healthy individuals. The results of this comparison were presented in an Excel-compatible, flat text format and then converted to RatDEGdb containing information about differential gene expression in the rat used as a model animal in biomedical research (URL=https://www.sysbio.ru/RatDEGdb). The conversion was performed using MariaDB 10.2.12, a freely available database (MariaDB Corp AB, Finland).

Likewise, Lu (2011) submitted a representative selection of PubMed publications telling about the current diversity of laboratory rat strains used as biomedical models simulating human diseases and about experimental methods to

Table 1. Primers for quantitative real-time polymerase chain reaction (qPCR)

Gene	Primers: 5′→3′	
	forward	reverse
Asmtl	CGCACTTCTCGGAGGTCCCGC	ACGGTCGCAGGGCTTCCCCA
Ppia	TTCCAGGATTCATGTGCCAG	CTTGCCATCCAGCCACTC
Rpl30	CATCTTGGCGTCTGATCTTG	TCAGAGTCTGTTTGTACCCC

Note. Primers were selected using the freely available web service PrimerBLAST (Ye et al., 2012). Rat genes: *Asmtl*, acetylserotonin O-methyltransferase like; *Ppia*, peptidylprolyl isomerase A; *Rpl30*, ribosomal protein L30. qPCR, quantitative real-time polymerase chain reaction using two reference genes as recommended by Bustin et al. (2009). The reference genes of our choice were *Ppia* (Gholami et al., 2017) and *Rpl30* (Penning et al., 2007) (for experimental substantiation, see our previous works (Chadaeva et al., 2021)).

assess differential gene expression with. Next, all rat DEGs in this selection of papers were documented and uploaded to RatDEGdb together with their supervised annotations, using an algorithm similar to the one described above for hypothalamic deficiency of *Asmtl* in aggressive rats. The lists of homologous genes were taken from the paralogs section of the GeneCards database (Stelzer et al., 2016). RatDEGdb includes the statistical significance of each DEG according to the estimates provided in the papers as referenced.

Statistical analysis of the differential expression of the *Asmtl* gene in the hypothalamus of the tame and aggressive rats used as an animal model of human aggressive behavior was performed using the menu "Statistics → Nonparametric → Mann–Whitney test" in STATISTICA (StatSoft™, USA), when two independent statistical criteria are being assessed at once: the nonparametric Mann–Whitney U test and the parametric test Fisher's Z, to assess the sustainability of results.

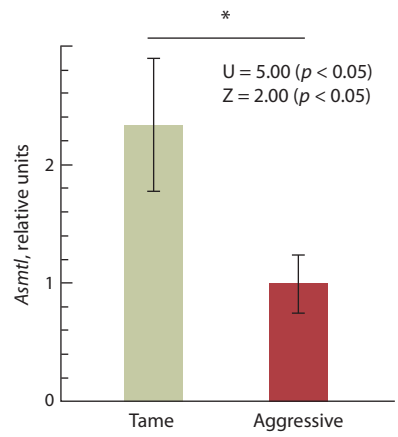


Fig. 1. Statistically significant differences in hypothalamic *Asmtl* expression levels between tame and aggressive adult male rats.

* Significance level $p < 0.05$ according to two independent statistical criteria: the nonparametric Mann–Whitney U test and the parametric test Fisher's Z, which reflects the sustainability of assessment results for *Asmtl* as a differentially expressed gene (DEG) in aggressive versus tame rats.

Results

Lower hypothalamic *Asmtl* mRNA levels in aggressive than in tame rats

Asmtl mRNA levels in the hypothalamus as measured and compared between the aggressive and tame rats are presented in Table 2. As can be seen from Figure 1, significantly lower *Asmtl* mRNA levels were in the aggressive than in tame rats in the settings of this experiment ($p < 0.05$; the Mann–Whitney U test and Fisher's Z).

Clinical manifestations of human *ASMTL* and *ASMT* deficiency

Table 3 presents the PubMed search results, with search terms (Lu, 2011) relating to human diseases associated with low expression levels of the *ASMTL* gene and its human paralog, *ASMT*. Line 1: the *Asmt*-deleted mouse models of human diseases (Trent et al., 2013) suggest a neurodevelopmental problem in the form of attention-deficit/hyperactivity disorder in combination with externalization symptoms (aggressive behavior) in children (Kang et al., 2023).

Line 2: *ASMT* deficiency is a molecular marker of autism, according to Melke and co-workers (2008), while a recent survey of teenagers above 12 years of age with autism spectrum disorders and epilepsy in past medical history revealed their inclination to aggression (Gaitanis et al., 2023).

These two examples are in favor of rather than against the low expression levels of the human genes *ASMTL* and *ASMT* representing, at least, combined molecular characteristics of the predisposition to some forms of aggressive behavior.

Finally, as can be seen from Table 3, these human genes were expressed at low levels among candidate molecular markers of a wide range of human diseases not associated with aggression: depression (Talarowska et al., 2014), developmental abnormalities (Li W. et al., 2012), brain injury (Govindarajulu et al., 2022; Yang et al., 2023), cell aging (Liu X. et al., 2022), cancer (Bi et al., 2019; Lau, Zhang, 2000; Xie et al., 2020; Cucielo et al., 2022; Liu Y. et al., 2022), infertility (Gonzalez-Arto et al., 2016; Zhang Z. et al., 2018) and asthma (Wu et al., 2020).

Put together, these findings reflect the fact that ASMT gene encoding the melatonin synthesis enzyme acetylserotonin O-methyltransferase is one of the key hormones involved in the regulation of molecular and genetic processes in all human body in general including aggression (Melke et al., 2008;

Table 2. Experimental data on "glove" test behavior and *Asmtl* mRNA levels for 12 adult male rats

Test	Strain	Outbred unrelated adult male tame and aggressive rats, hypothalamus						$M_0 \pm SEM$
		# 1	# 2	# 3	# 4	# 5	# 6	
"Glove" test	A	–3	–3	–3	–3	–3	–3	
	T	3	3	3	3	3	3	
qPCR (*Asmtl*)	A	1.88 ± 0.67	0.80 ± 1.65	1.56 ± 0.51	0.70 ± 0.04	0.33 ± 0.16	0.73 ± 0.02	1.00 ± 0.24
	T	4.51 ± 0.51	1.21 ± 0.15	1.73 ± 0.63	0.92 ± 0.04	3.30 ± 0.09	2.33 ± 0.13	2.33 ± 0.56

Note. see Notes to Table 1. Rat strain: A, aggressive rats ($n = 6$); T, tame rats ($n = 6$). Tests: "glove" test, in which each rat was scored from "–4" (most aggressive) to "+4" (most friendly), according to a work by Plyusnina and Oskina (1997); *Asmtl* expression levels, $M_0 \pm SEM$, estimates of the mean ± standard error of the mean from three technical replicates, with a LightCycler® 96 operated in the automatic mode (Roche, Switzerland).

Table 3. Clinical manifestation of deficiencies in *ASMTL* and in its human paralogue *ASMT* in human diseases according to the current release of the RatDEGdb knowledge base

No.	Disease	Clinical manifestation of ASMTL and ASMT deficiency	References
1	Neurodevelopmental disorders	In *Asmt*-deleted mice used as models of human diseases: neurodevelopmental disorders, attention-deficit/hyperactivity disorder	Trent et al., 2013
2	Autism	In a cohort study: low *ASMT* mRNA levels in the blood and low *ASMT* gene expression leading to melatonin deficiency may be molecular markers of autism	Melke et al., 2008
3	Depression with speech and learning disorders	In a cohort clinical study: *ASMT* deficiency as a marker of recurrent depressive disorder with impaired speech fluency and auditory-verbal learning	Talarowska et al., 2014
4	Depression with sleep and circadian rhythm disorders	In human behavioral models using *Asmt*-knockout mice: depression, sleep and circadian rhythm disturbances that altogether may be reversed due to swimming exercise	Liu W. et al., 2022
5	Recovery from acute traumatic brain injury (concussion)	In models of acute human brain injury using rats exposed to strong sensory stimuli: decrease in *Asmt* levels after 6 hours, 24 hours and even 1 month after exposure – manifestations of brain contusion in the form of sleep disturbances	Govindarajulu et al., 2022
6	Cerebral hypoxia-ischemia	In human disease models using neonatal rats: *Asmt* deficiency may be a molecular marker of cerebral hypoxia-ischemia	Yang et al., 2023
7	Developmental disorders	In human disease models using induced pluripotent stem cell lines derived from skin fibroblasts from patients with any developmental disorder: *ASMTL* deficiency may be one of the most common molecular markers of developmental disorders	Li W. et al., 2012
8	Cellular senescence	In human ageing models using cell cultures: slowing down replicative senescence of human bone marrow mesenchymal stromal cells	Liu X. et al., 2022
9	Glioma	In a retrospective transcriptome meta-analysis summarizing 966 glioma-related RNA-seq and microarray assay dataset: *ASMT* deficiency may be a clinical molecular marker of glioma	Liu Y. et al., 2022
10	Colon cancer	In human disease models using colon cancer cell lines LOVO and HCT116: cancer cell proliferation, migration and invasion decreased with downregulation of *ASMTL* expression	Bi et al., 2019
11	Prostate cancer	In a cohort study of patients using qPCR technology: *ASMTL*-upregulation promotes the development of prostate cancer	Lau, Zhang, 2000
12	Ovarian cancer	In a clinical cohort study: *ASMT*-deficient patients with ovarian cancer had a decrease in median survival by several months	Cucielo et al., 2022
13	Breast cancer	In a cohort study: *ASMT*-inhibitors reduce the invasiveness of breast cancer cells	Xie et al., 2020
14	Subfertility	In human fertility models using rams: reduced sperm capacitation; selection of a line of laboratory mice with a functional *Asmt* allele: most lines of laboratory mice have dysfunction of this gene, due to which melatonin deficiency in them reduces its negative impact on their spermatogenesis	Gonzalez-Arto et al., 2016; Zhang Z. et al., 2018
15	Respiratory tract inflammation, asthma	In mouse models of human disease: *Asmt* deficiency promotes airway inflammation such as asthma due to melatonin deficiency	Wu et al., 2020
Total	19 diseases	24 clinical manifestations of *ASMTL* or *ASMT* deficiency	16 references

Trent et al., 2013; Gaitanis et al., 2023; Kang et al., 2023), depression (Talarowska et al., 2014), ontogenesis (Li W. et al., 2012; Zhang Z. et al., 2018), wound healing (Govindarajulu et al., 2022; Yang et al., 2023), ageing (Liu X. et al., 2022) oncoprotector (Lv et al., 2019), anti-inflammatory and immunomediatory mechanisms (Li G. et al., 2021).

RatDEGdb: the knowledge base

Figure 2 shows how RatDEGdb compares the hypothalamic level of *Asmtl* in the aggressive rat strain with that in the tame. Here aggression is considered to be a comorbid symptom in human diseases such as thalassemia, obesity and carcinoma (for review, see Chadaeva et al., 2016). Consequently,

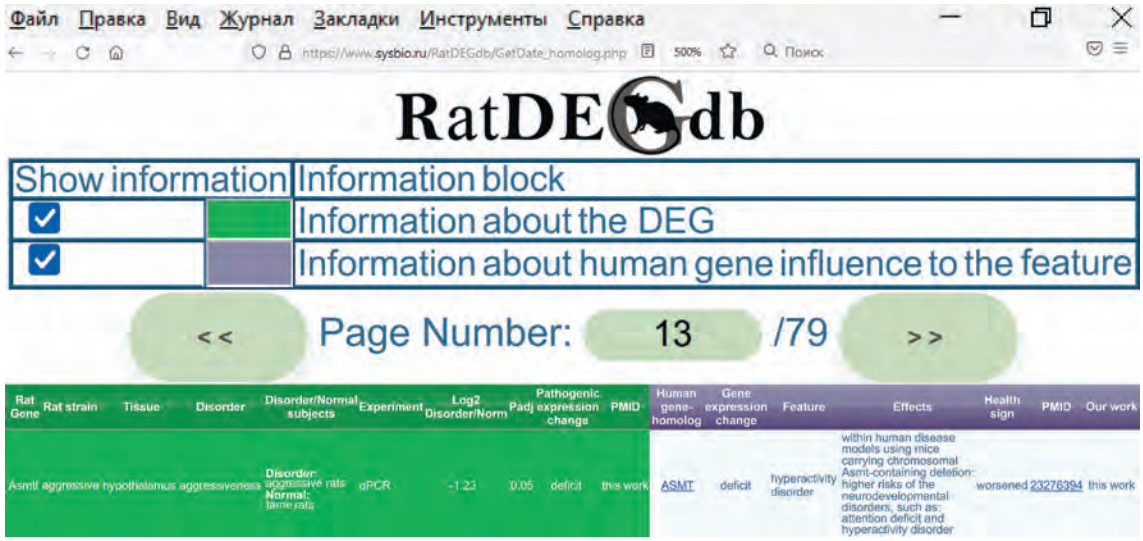


Fig. 2. A sample entry in RatDEGdb documents original experimental data on *Asmtl* deficiency in the hypothalamus of aggressive rats compared to the tame rats as a biomedical model of aggressive behavior in human diseases (see Fig. 1 and Table 2) together with their annotation (see Table 3: first row) using independent data on low expression levels of its human homolog *ASMT* in patients with hyperactivity disorders according to an *Asmt*-deleted mouse model of human disease (Trent et al., 2013).

RatDEGdb (see Fig. 1 and Table 2) integrated data on low hypothalamic levels of the *Asmtl* gene in the aggressive rats and low levels of its human homolog *ASMTL* as found in patients with neurodevelopmental problems in the form of attention-deficit/hyperactivity disorder using an *Asmt*-deleted mouse model (Trent et al., 2013) (see Table 3).

The current release contains information on DEGs in ten genetic rat strains used as models of 11 human pathologies (Tables 4–6). As can be seen in the bottom lines of these tables, RatDEGdb now contains information on 25,101 DEGs representing 14,320 unique rat genes that change transcription levels in 21 tissues of 10 genetic rat strains used as models of 11 human diseases based on 45 original scientific papers referenced in the rightmost column of Tables 4–6. These rat DEGs were annotated with information about equal changes in the expression levels of their human homologs in affected people. In total, the current release contains 94,873 such annotations for 321 human genes in 836 diseases based on 959 PubMed publications (Lu, 2011). Thus, RatDEGdb is unique in that the manual curation of the annotation of DEGs of the rat as a model object simulating human pathology uses independent clinical data, which none of other biomedical databases does.

Discussion

The elementary step in filling RatDEGdb with data can be seen in Tables 1–3 and Figures 1–2, with the *Asmtl* (acetylserotonin O-methyltransferase like) gene as an example. The hypothalamic expression of this gene was profiled and compared between aggressive and tame rats used as model animals in human aggression research. Results of the analysis of this gene by real-time PCR are provided. These results were annotated using PubMed papers (Lu, 2011) about equal changes in the expression levels of its human homologs *ASMTL* and *ASMT* in patients. Then this annotation of the *Asmtl* gene differentially expressed in the hypothalamus of the aggressive and tame rats was supplemented with PCR-, RNA-seq- and microarray-based information on all DEGs in the rat used as a model object in biomedical research. Next, the uncharacterized, unannotated, predicted, and not protein-encoding genes were dropped. Finally, we annotated the remaining rat DEGs with publicly available works about the clinical manifestations of equal changes in the expression levels of their human homologs in patients, put these annotations together as the RatDEGdb the knowledge base, and made it freely available at https://www.sysbio.ru/RatDEGdb.

Figures 1 and 2 show how RatDEGdb characterizes the DEGs of various breeding-based rat strains primarily developed in the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia). The ISIAH rats were used as model animals in the biomedical studies of stress-induced arterial hypertension, as summarized in Tables 4 and 5. The same tables show that that OXYS rats were used for studying age-related diseases and ageing processes; and GC rats, for studying psychopathological conditions (see Table 4). In addition, tame and aggressive rat strains were used for studying animal domestication (Plyusnina, Oskina, 1997; Gulevich et al., 2019; Chadaeva et al., 2021) and aggression (Popova et al., 2010) as symptoms of obesity and thalassemia (Chadaeva et al., 2016, 2019). As can seen from Tables 4–6, whole-genome sequencing was performed on each of these models, except for the GC strain, in which only the expression levels of the glutamate receptor genes and the catecholamine system genes were measured.

The existing biomedical databases intended for studying human diseases are normally focused on the information on the human genome (Stenson et al., 2014; Singh et al., 2018;

Table 4. Characterization of the qPCR-inferred DEGs of the rat as a model animal in biomedicine documented in the RatDEGdb knowledge base

#	Strain	Tissue	Disease	Model	Norm	N_{DEG}	References
1	Aggressive	hyp	Aggression	Aggressive	Tame	1	This work
2		hyp	Aggression	Aggressive	Tame	4	Klimova et al., 2021
3		fc, hip, hyp, mb	Aggression	Aggressive	Tame	21	Moskaliuk et al., 2023
4		hip, hyp, mb	Aggression	Aggressive	Tame	11	Moskaliuk et al., 2022
5		hyp	Aggression	Aggressive	Tame	8	Klimova et al., 2021
6		hyp	Aggression	Aggressive	Tame	3	Gulevich et al., 2019
7		mb, hip, fc	Aggression	Aggressive	Tame	5	Kondaurova et al., 2016
8		hip, hyp, mb	Aggression	Aggressive	Tame	3	Ilchibaeva et al., 2016
9		hyp	Aggression	Aggressive	Tame	7	Oshchepkov et al., 2019
10		hip, hyp, mb	Aggression	Aggressive	Tame	7	Ilchibaeva et al., 2015
11		fc, hip	Aggression	Aggressive	Tame	2	Popova et al., 2010
12		fc	Aggression	Aggressive	Tame	1	Naumenko et al., 2009
13		mb	Aggression	Aggressive	Tame	1	Popova et al., 2007
14		hip	Aggression	Aggressive	Tame	4	Herbeck et al., 2010
15	Tame	hip	Aggression	Tame, methil	Tame	3	Herbeck et al., 2010
16	SD	mpc, ac, pc, ic	Aggression	Isolation	Socialized	22	Wall et al., 2012
17		Brain	Aggression	Aggressive	Non-aggress	5	Suzuki et al., 2010
18		hip	Autism	SD, PPA	SD	6	Choi et al., 2018
19	GC	hip	Catatonia	GC	WAG	1	Plekanchuk, Ryazanova, 2021
20		mb	Catatonia	GC	WAG	1	Ryazanova et al., 2017
21	ISIAH	Kidney, myoc	HT	ISIAH	WAG	6	Fedoseeva et al., 2011
22		hyp, mo	HT	ISIAH	WAG	3	Klimov et al., 2013
23		Kidney	HT	ISIAH	WAG	1	Fedoseeva et al., 2009
24		hyp, mo	HT	ISIAH, $	ISIAH	5	Klimov et al., 2017
25	OXYS	Retina	AMD	OXYS	Wistar	5	Perepechaeva et al., 2014
26		Retina	AMD	OXYS, SkQ1	OXYS	5	Perepechaeva et al., 2014
27		Retina	AMD	Wistar, SkQ1	Wistar	2	Perepechaeva et al., 2014
Total	6 strains	14 tissues	5 diseases	11 models	7 models	143	23 references

Note. Here and in Tables 5 and 6 : N_{DEG}, number of DEGs. Tissues: ac, anterior cingulate; ag, adrenal gland; bmmscs, bone marrow-derived mesenchymal stromal cells; bmp, brain microvascular pericytes; bs, brain stem; fc, frontal cortex; hip, hippocampus; hyp, hypothalamus; ic, infralimbic cortex; lvcp, lateral ventricular choroid plexus; mb, midbrain; mo, medulla oblongata; mpc, medial prefrontal cortex; mt, midbrain tegmentum; myoc, myocardium; PAG, periaqueductal gray matter; pc, prelimbic cortex; po, prefrontal cortex; rc, renal cortex; rm, renal medulla. Diseases: AD, Alzheimer's disease; AMD, age-related macular degeneration; ARBLBD, age-related blood-liquor barrier development; CRS, cellular replicative senescence; HT, hypertension; PAH, pulmonary arterial hypertension.
Models: $, Agtr1a-blocker; PPA, propionic acid; SkQ1, Skulachev's antioxidant.

Sun et al., 2022) and contain primary transcriptome information. RatDEGdb is novel in that it supplements biomedicine-based whole-genome experimental data on rat DEGs with clinical data on equal changes in the expression levels of their human homologs in patients, for further use of all these data in personalized medicine. With a new capability that enables the researcher to compare pathogenic changes in gene expression in humans and model animals, RatDEGdb can be useful in addressing problems in systems biology and clinical medicine.

Table 5. Characterization of the RNA-seq-inferred DEGs of the rat as a model animal in biomedicine documented in the RatDEGdb knowledge base

#	Strain	Tissue	Disease	Model	Norm	N_{DEG}	References
1	Aggressive	fc	Aggression	Aggressive	Tame	24	Albert et al., 2012
2		hyp	Aggression	Aggressive	Tame	46	Chadaeva et al., 2021
3		hip	Aggression	Aggressive	Tame	42	Oshchepkov et al., 2022a
4		mt	Aggression	Aggressive	Tame	31	Oshchepkov et al., 2022b
5		PAG	Aggression	Aggressive	Tame	39	Shikhevich et al., 2023
6	ISIAH	bs	HT	ISIAH	WAG	206	Fedoseeva et al., 2019
7		hyp	HT	ISIAH	WAG	137	Klimov et al., 2016
8		rm	HT	ISIAH	WAG	882	Ryazanova et al., 2016
9		rc	HT	ISIAH	WAG	309	Fedoseeva et al., 2016b
10		ag	HT	ISIAH	WAG	1020	Fedoseeva et al., 2016a
11	OXYS	hip	AD	OXYS, 20 do	Wistar, 20 do	46	Stefanova et al., 2018
12		hip	AD	OXYS, 5 mo	Wistar, 5 mo	28	Stefanova et al., 2018
13		hip	AD	OXYS, 18 mo	Wistar, 18 mo	85	Stefanova et al., 2018
14		po	AD	OXYS, 20 do	Wistar, 20 do	2	Stefanova et al., 2019
15		po	AD	OXYS, 5 mo	Wistar, 5 mo	7	Stefanova et al., 2019
16		po	AD	OXYS, 18 mo	Wistar, 18 mo	73	Stefanova et al., 2019
17		Retina	AMD	OXYS, 3 mo	Wistar, 3 mo	117	Kozhevnikova et al., 2013
18		Retina	AMD	OXYS, 18 mo	Wistar, 18 mo	85	Kozhevnikova et al., 2013
19		po	AD	OXYS, 5 mo	OXYS, 20 do	52	Stefanova et al., 2019
20		po	AD	OXYS, 18 mo	OXYS, 5 mo	58	Stefanova et al., 2019
21		hip	AD	OXYS, 5 mo	OXYS, 20 do	135	Stefanova et al., 2018
22		hip	AD	OXYS, 18 mo	OXYS, 5 mo	197	Stefanova et al., 2018
23		Retina	AMD	OXYS, 18 mo	OXYS, 3 mo	19	Kozhevnikova et al., 2013
24	Wistar	hip	AD	Wistar, 5 mo	Wistar, 20 do	150	Stefanova et al., 2018
25		hip	AD	Wistar, 18 mo	Wistar, 5 mo	190	Stefanova et al., 2018
26		Retina	AMD	Wistar, 18 mo	Wistar, 3 mo	28	Kozhevnikova et al., 2013
25	SD	bmmscs	CRS	SD, 20 p	SD, 5 p	9167	Liu X. et al., 2022
26		bmmscs	CRS	SD, 5 p	SD, 5 p, ASA	1220	Liu X. et al., 2022
27		bmmscs	CRS	SD, 20 p	SD, 20 p, ASA	446	Liu X. et al., 2022
28		lvcp	ARBLBD	SD, 6 wo	SD, 15 ed	9159	Liddelow et al., 2013
29		Lung	PAH	SD, MCT	SD	40	Xiao et al., 2020
30		rc	HT	SD, l-NAME	SD	284	Tain et al., 2015
31		rc	HT	SD, DEX	SD	44	Tain et al., 2015
32		rc	HT	SD, hfd	SD	240	Tain et al., 2015
33	SHR	Kidney	HT	SHR	WKY	68	Watanabe et al., 2015
34		bmp	HT	SHR	WKY	21	Yuan et al., 2018
35	SHRSP	Kidney	Stroke	SHRSP	WKY	27	Watanabe et al., 2015
36	DSS	Kidney	HT	DSS	DSS, QSYQ	13	Du et al., 2021
37		Kidney	HT	DSS, $Resp18^{MUT}$	DSS	14	Ashraf et al., 2021
Total	8 strains	17 tissues	8 diseases	17 models	17 models	24751	21 references

Note. Models: ASA, aspirin; do, days old; DEX, dexamethasone; ed, embryonic days; hfd, high-fructose diet; l-NAME, NG-nitro-l-arginine-methyester; MCT, monocrotaline; $Resp18^{MUT}$, mutant variant; mo, months old; p, passage old; QSYQ, Chinese traditional medicine prescription Qi-Shen-Yi-Qi; wo, weeks old.

Table 6. Characterization of the microarray-inferred DEGs of the rat as a model animal in biomedicine documented in the RatDEGdb knowledge base

#	Strain	Tissue	Disease	Model	Norm	N_{DEG}	References
1	Wistar	ag	HT	Wistar, DEX	Wistar	93	Tharmalingam et al., 2020
2	SHR	ag	HT	SHR, 3 wo	WKY, 3 wo	12	Yoshida et al., 2014
3		ag	HT	SHR, 6 wo	WKY, 6 wo	42	Yoshida et al., 2014
4		Brain	HT	SHR, 3 wo	WKY, 3 wo	11	Yoshida et al., 2014
5		Brain	HT	SHR, 6 wo	WKY, 6 wo	10	Yoshida et al., 2014
6	SHRSP	ag	Stroke	SHR, 3 wo	SHR, 3 wo	17	Yoshida et al., 2014
7		ag	Stroke	SHR, 6 wo	SHR, 6 wo	9	Yoshida et al., 2014
8		Brain	Stroke	SHR, 6 wo	SHR, 6 wo	11	Yoshida et al., 2014
9		Brain	Stroke	SHR, 3 wo	SHR, 3 wo	2	Yoshida et al., 2014
Total	3 strains	2 tissues	2 diseases	3 models	5 models	207	2 references

Note. Models: wo, weeks old.

Conclusion

The RatDEGdb knowledge base is a collection of experimental data and a toolkit for interactive analyses in genomic research into diseases, such as Alzheimer's disease, autism, hypertension and some others. We are planning to continue updating RatDEGdb by adding new information on gene expression in rats as model objects of human diseases and annotating the DEGs with pieces of works on equal changes in the expression levels of their human homologs in patients.

References

Aikawa H., Nonaka I., Woo M., Tsugane T., Esaki K. Shaking rat Kawasaki (SRK): a new neurological mutant rat in the Wistar strain. *Acta Neuropathol*. 1988;76:366-372. DOI 10.1007/BF00686973

Albert F.W., Somel M., Carneiro M., Aximu-Petri A., Halbwax M., Thalmann O., Blanco-Aguiar J.A., Plyusnina I.Z., Trut L., Villafuerte R., Ferrand N., Kaiser S., Jensen P., Paabo S. A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet*. 2012;8(9):e1002962. DOI 10.1371/journal.pgen.1002962

Ashraf U.M., Mell B., Jose P.A., Kumarasamy S. Deep transcriptomic profiling of Dahl salt-sensitive rat kidneys with mutant form of *Resp18*. *Biochem. Biophys. Res. Commun*. 2021;572:35-40. DOI 10.1016/j.bbrc.2021.07.071

Barykina N.N., Chepkasov I.L., Alekhina T.A., Kolpakov V.G. Selection of Wistar rats for predisposition to catalepsy. *Genetika*. 1983;19(12):2014-2021

Bay V., Happ D.F., Ardalan M., Quist A., Oggiano F., Chumak T., Hansen K., Ding M., Mallard C., Tasker R.A., Wegener G. Flinders sensitive line rats are resistant to infarction following transient occlusion of the middle cerebral artery. *Brain Res*. 2020;1737:146797. DOI 10.1016/j.brainres.2020.146797

Belyaev D.K., Borodin P.M. The influence of stress on variation and its role in evolution. *Biologisches Zentralblatt*. 1982;101(6):705-714

Bi J., Huang Y., Liu Y. Effect of NOP2 knockdown on colon cancer cell proliferation, migration, and invasion. *Transl. Cancer Res*. 2019;8(6):2274-2283. DOI 10.21037/tcr.2019.09.46

Bustin S.A., Benes V., Garson J.A., Hellemans J., Huggett J., Kubista M., Mueller R., Nolan T., Pfaffl M.W., Shipley G.L., Vandesompele J., Wittwer C.T. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem*. 2009;55(4):611-622. DOI 10.1373/clinchem.2008.112797

Carter C.S., Richardson A., Huffman D.M., Austad S. Bring back the rat! *J. Gerontol. A Biol. Sci. Med. Sci*. 2020;75(3):405-415. DOI 10.1093/gerona/glz298

Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A., Sharypova E.B., Kashina E.V., Matveeva M.Y., Arshinova T.V., Ponomarenko P.M., Arkova O.V., Bondar N.P., Savinkova L.K., Kolchanov N.A. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics*. 2016;17(Suppl.14):995. DOI 10.1186/s12864-016-3353-3

Chadaeva I., Ponomarenko P., Rasskazov D., Sharypova E., Kashina E., Kleshchev M., Ponomarenko M., Naumenko V., Savinkova L., Kolchanov N., Osadchuk L., Osadchuk A. Natural selection equally supports the human tendencies in subordination and domination: a genome-wide study with *in silico* confirmation and *in vivo* validation in mice. *Front. Genet*. 2019;10:73. DOI 10.3389/fgene.2019.00073

Chadaeva I., Ponomarenko P., Kozhemyakina R., Suslov V., Bogomolov A., Klimova N., Shikhevich S., Savinkova L., Oshchepkov D., Kolchanov N., Markel A., Ponomarenko M. Domestication explains two-thirds of differential-gene-expression variance between domestic and wild animals; the remaining one-third reflects intraspecific and interspecific variation. *Animals*. 2021;11(9):2667. DOI 10.3390/ani11092667

Choi J., Lee S., Won J., Jin Y., Hong Y., Hur T.Y., Kim J.H., Lee S.R., Hong Y. Pathophysiological and neurobehavioral characteristics of a propionic acid-mediated autism-like rat model. *PLoS One*. 2018;13(2):e0192925. DOI 10.1371/journal.pone.0192925

Cucielo M.S., Cesario R.C., Silveira H.S., Gaiotte L.B., Dos Santos S.A.A., de Campos Zuccari D.A.P., Seiva F.R.F., Reiter R.J., de Almeida Chuffa L.G. Melatonin reverses the warburg-type metabolism and reduces mitochondrial membrane potential of ovarian cancer cells independent of MT1 receptor activation. *Molecules*. 2022;27(14):4350. DOI 10.3390/molecules27144350

Du H., Xiao G., Xue Z., Li Z., He S., Du X., Zhou Z., Cao L., Wang Y., Yang J., Wang X., Zhu Y. QiShenYiQi ameliorates salt-induced hypertensive nephropathy by balancing ADRA1D and SIK1 expression in Dahl salt-sensitive rats. *Biomed. Pharmacother*. 2021;141: 111941. DOI 10.1016/j.biopha.2021.111941

Fedoseeva L.A., Dymshits G.M., Markel A.L., Jakobson G.S. Renin system of the kidney in ISIAH rats with inherited stress-induced arterial hypertension. *Bull. Exp. Biol. Med*. 2009;147(2):177-180. DOI 10.1007/s10517-009-0465-7

Fedoseeva L.A., Riazanova M.A., Antonov E.V., Dymshits G.M., Markel A.L. Expression of the renin angiotensin system genes in the kidney and heart of ISIAH hypertensive rats. *Biochem. Moscow Suppl. Ser. B*. 2011;5(1):37-43. DOI 10.1134/s1990750811010069

Fedoseeva L.A., Klimov L.O., Ershov N.I., Alexandrovich Y.V., Efimov V.M., Markel A.L., Redina O.E. Molecular determinants of the adrenal gland functioning related to stress-sensitive hypertension in ISIAH rats. *BMC Genomics*. 2016a;17(Suppl.14):989. DOI 10.1186/s12864-016-3354-2

Fedoseeva L.A., Ryazanova M.A., Ershov N.I., Markel A.L., Redina O.E. Comparative transcriptional profiling of renal cortex in rats with inherited stress-induced arterial hypertension and normotensive Wistar Albino Glaxo rats. *BMC Genet*. 2016b;17(Suppl.1):12. DOI 10.1186/s12863-015-0306-9

Fedoseeva L.A., Klimov L.O., Ershov N.I., Efimov V.M., Markel A.L., Orlov Y.L., Redina O.E. The differences in brain stem transcriptional profiling in hypertensive ISIAH and normotensive WAG rats. *BMC Genomics*. 2019;20(Suppl.3):297. DOI 10.1186/s12864-019-5540-5

Gaitanis J., Nie D., Hou T., Frye R. Developmental regression followed by epilepsy and aggression: a new syndrome in autism spectrum disorder? *J. Pers. Med*. 2023;13(7):1049. DOI 10.3390/jpm 13071049

Gayday E.A., Gayday D.S. Genetic diversity of experimental mice and rats: history of origin, methods of production and check. *Laboratornye Zhivotnye Dlya Nauchnykh Issledovaniy = Laboratory Animals for Scientific Research*. 2019;4:78-85. DOI 10.29296/26187 23X-2019-04-09 (in Russian)

Gholami K., Loh S.Y., Salleh N., Lam S.K., Hoe S.Z. Selection of suitable endogenous reference genes for qPCR in kidney and hypothalamus of rats under testosterone influence. *PLoS One*. 2017;12(6): e0176368. DOI 10.1371/journal.pone.0176368

Gibbs R.A., Weinstock G.M., Metzker M.L., Muzny D.M., Sodergren E.J., Scherer S., Scott G., Steffen D., Worley K.C., Burch P.E., … Peterson J., Guyer M., Felsenfeld A., Old S., Mockrin S., Collins F; Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004;428(6982):493-521. DOI 10.1038/nature02426

Gonzalez-Arto M., Hamilton T.R., Gallego M., Gaspar-Torrubia E., Aguilar D., Serrano-Blesa E., Abecia J.A., Perez-Pe R., Muino-Blanco T., Cebrian-Perez J.A., Casao A. Evidence of melatonin synthesis in the ram reproductive tract. *Andrology*. 2016;4(1):163-171. DOI 10.1111/andr.12117

Govindarajulu M., Patel M.Y., Wilder D.M., Long J.B., Arun P. Blast exposure dysregulates nighttime melatonin synthesis and signaling in the pineal gland: a potential mechanism of blast-induced sleep disruptions. *Brain Sci*. 2022;12(10):1340. DOI 10.3390/brainsci 12101340

Greenhouse D.D., Festing M.F.W., Hasan S., Cohen A.L. Inbred strains of rats and mutants. In: Hedrich H.J. (Ed.) Genetic Monitoring of Inbred Strains of Rats. Stuttgart: Gustav Fischer Verlag, 1990; 410-480

Gryksa K., Schmidtner A.K., Masís-Calvo M., Rodríguez-Villagra O.A., Havasi A., Wirobski G., Maloumby R., Jägle H., Bosch O.J., Slattery D.A., Neumann I.D. Selective breeding of rats for high (HAB) and low (LAB) anxiety-related behaviour: a unique model for comorbid depression and social dysfunctions. *Neurosci. Biobehav. Rev*. 2023;152:105292. DOI 10.1016/j.neubiorev.2023.105292

Gulevich R., Kozhemyakina R., Shikhevich S., Konoshenko M., Herbeck Y. Aggressive behavior and stress response after oxytocin administration in male Norway rats selected for different attitudes to humans. *Physiol. Behav*. 2019;199:210-218. DOI 10.1016/j.physbeh. 2018.11.030

Herbeck Yu.E., Os'kina I.N., Gulevich R.G., Plyusnina I.Z. Effects of maternal methyl-supplemented diet on hippocampal glucocorticoid receptor mRNA expression in rats selected for behavior. *Cytol. Genet.* (*Moscow.*). 2010;44(2):108-113. DOI 10.3103/S00954527 10020064

Ideno J., Mizukami H., Honda K., Okada T., Hanazono Y., Kume A., Saito T., Ishibashi S., Ozawa K. Persistent phenotypic correction of central diabetes insipidus using adeno-associated virus vector expressing arginine-vasopressin in Brattleboro rats. *Mol. Ther*. 2003; 8(6):895-902. DOI 10.1016/j.ymthe.2003.08.019

Ilchibaeva T.V., Kondaurova E.M., Tsybko A.S., Kozhemyakina R.V., Popova N.K., Naumenko V.S. Brain-derived neurotrophic factor (BDNF) and its precursor (proBDNF) in genetically defined fear-induced aggression. *Behav. Brain Res*. 2015;290:45-50. DOI 10.1016/ j.bbr.2015.04.041

Ilchibaeva T.V., Tsybko A.S., Kozhemyakina R.V., Naumenko V.S. Expression of apoptosis genes in the brain of rats with genetically defined fear-induced aggression. *Mol. Biol.* (*Moscow*). 2016;50(5): 814-820. DOI 10.7868/S0026898416030071

Kang S., Gair S.L., Paton M.J., Harvey E.A. Racial and ethnic differences in the relation between parenting and preschoolers' externalizing behaviors. *Early Educ. Dev*. 2023;34(4):823-841. DOI 10.1080/10409289.2022.2074202

Klimov L.O., Fedoseeva L.A., Ryazanova M.A., Dymshits G.M., Markel A.L. Expression of renin-angiotensin system genes in brain structures of ISIAH rats with stress-induced arterial hypertension. *Bull. Exp. Biol. Med*. 2013;154(3):357-660. DOI 10.1007/s10517-013-1950-6

Klimov L.O., Ershov N.I., Efimov V.M., Markel A.L., Redina O.E. Genome-wide transcriptome analysis of hypothalamus in rats with inherited stress-induced arterial hypertension. *BMC Genet*. 2016; 17(Suppl.1):13. DOI 10.1186/s12863-015-0307-8

Klimov L.O., Ryazanova M.A., Fedoseeva L.A., Markel A.L. Effects of brain renin-angiotensin system inhibition in ISIAH rats with inherited stress-induced arterial hypertension. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2017; 21(6):735-741. DOI 10.18699/VJ17.29-o (in Russian)

Klimova N.V., Chadaeva I.V., Shichevich S.G., Kozhemyakina R.V. Differential expression of 10 genes in the hypothalamus of two generations of rats selected for a reaction to humans. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(2):208-215. DOI 10.18699/VJ21.50-o

Kolosova N.G., Stefanova N.A., Korbolina E.E., Fursova A.Z., Kozhevnikova O.S. Senescence-accelerated OXYS rats: a genetic model of premature aging and age-related diseases. *Adv. Gerontol*. 2014;4:294-298. DOI 10.1134/S2079057014040146

Kolpakov V.G., Kulikov A.V., Alekhina T.A., Chuguy V.F., Petrenko O.I., Barykina N.N. Catatonia or depression: the GC rat strain as an animal model of psychopathology. *Russ. J. Genet*. 2004;40(6): 672-678. DOI 10.1023/B:RUGE.0000033315.79449.d4

Kondaurova E.M., Ilchibaeva T.V., Tsybko A.S., Kozhemyakina R.V., Popova N.K., Naumenko V.S. 5-HT1A receptor gene silencers Freud-1 and Freud-2 are differently expressed in the brain of rats with genetically determined high level of fear-induced aggression

or its absence. *Behav. Brain Res*. 2016;310:20-25. DOI 10.1016/j.bbr.2016.04.050

Kozhevnikova O.S., Korbolina E.E., Ershov N.I., Kolosova N.G. Rat retinal transcriptome: effects of aging and AMD-like retinopathy. *Cell Cycle*. 2013;12(11):1745-1761. DOI 10.4161/cc.24825

Lau Y.F., Zhang J. Expression analysis of thirty one Y chromosome genes in human prostate cancer. *Mol. Carcinog*. 2000;27(4):308-321. DOI 10.1002/(sici)1098-2744(200004)27:4<308::aid-mc9>3.0.co;2-r

Li G., Lv D., Yao Y., Wu H., Wang J., Deng S., Song Y., Guan S., Wang L., Ma W., Yang H., Yan L., Zhang J., Ji P., Zhang L., Lian Z., Liu G. Overexpression of ASMT likely enhances the resistance of transgenic sheep to brucellosis by influencing immune-related signaling pathways and gut microbiota. *FASEB J*. 2021;35(9):e21783. DOI 10.1096/fj.202100651r

Li W., Wang X., Fan W., Zhao P., Chan Y.C., Chen S., Zhang S., Guo X., Zhang Y., Li Y., Cai J., Qin D., Li X., Yang J., Peng T., Zychlinski D., Hoffmann D., Zhang R., Deng K., Ng K.M., Menten B., Zhong M., Wu J., Li Z., Chen Y., Schambach A., Tse H.F., Pei D., Esteban M.A. Modeling abnormal early development with induced pluripotent stem cells from aneuploid syndromes. *Hum. Mol. Genet*. 2012;21(1):32-45. DOI 10.1093/hmg/ddr435

Liddelow S.A., Dziegielewska K.M., Ek C.J., Habgood M.D., Bauer H., Bauer H.C., Lindsay H., Wakefield M.J., Strazielle N., Kratzer I., Mollgard K., Ghersi-Egea J.F., Saunders N.R. Mechanisms that determine the internal environment of the developing brain: a transcriptomic, functional and ultrastructural approach. *PLoS One*. 2013;8(7):e65629. DOI 10.1371/journal.pone.0065629

Liu W., Huang Z., Xia J., Cui Z., Li L., Qi Z., Liu W. Gene expression profile associated with Asmt knockout-induced depression-like behaviors and exercise effects in mouse hypothalamus. *Biosci. Rep*. 2022;42(7):bsr20220800. DOI 10.1042/bsr20220800

Liu X., Zhan Y., Xu W., Liu L., Liu X., Da J., Zhang K., Zhang X., Wang J., Liu Z., Jin H., Zhang B., Li Y. Characterization of transcriptional landscape in bone marrow-derived mesenchymal stromal cells treated with aspirin by RNA-seq. *PeerJ*. 2022;10:e12819. DOI 10.7717/peerj.12819

Liu Y., Xiang J., Liao Y., Peng G., Shen C. Identification of tryptophan metabolic gene-related subtypes, development of prognostic models, and characterization of tumor microenvironment infiltration in gliomas. *Front. Mol. Neurosci*. 2022;15:1037835. DOI 10.3389/fnmol.2022.1037835

Lu Z. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. Database (Oxford). 2011;2011:baq036. DOI 10.1093/database/baq036

Lv J.W., Zheng Z.Q., Wang Z.X., Zhou G.Q., Chen L., Mao Y.P., Lin A.H., Reiter R.J., Ma J., Chen Y.P., Sun Y. Pan-cancer genomic analyses reveal prognostic and immunogenic features of the tumor melatonergic microenvironment across 14 solid cancer types. *J. Pineal Res*. 2019;66(3):e12557. DOI 10.1111/jpi.12557

Markel A.L. Development of a new strain of rats with inherited stress-induced arterial hypertension. In: Sassard J. (Ed.) Genetic Hypertension. London: John Libbey Eurotext Ltd., 1992;218:405-407

Markel A.L., Maslova L.N., Shishkina G.T., Mahanova N.A., Jacobson G.S. Developmental influences on blood pressure regulation in ISIAH rats. In: McCarty R., Blizard D.A., Chevalier R.L. (Eds.) Development of the Hypertensive Phenotype: Basic and Clinical Studies. In the series Handbook of Hypertension. Amsterdam: Elsevier, 1999;493-526

Martín-Carro B., Donate-Correa J., Fernández-Villabrille S., Martín-Vírgala J., Panizo S., Carrillo-López N., Martínez-Arias L., Navarro-González J.F., Naves-Díaz M., Fernández-Martín J.L., Alonso-Montes C., Cannata-Andía J.B. Experimental models to study diabetes mellitus and its complications: limitations and new opportunities. *Int. J. Mol. Sci*. 2023;24(12):10309. DOI 10.3390/ijms241210309

Melke J., Goubran Botros H., Chaste P., Betancur C., Nygren G., Anckarsäter H., Rastam M., Ståhlberg O., Gillberg I.C., Delorme R., Chabane N., Mouren-Simeoni M.C., Fauchereau F., Durand C.M., Chevalier F., Drouot X., Collet C., Launay J.M., Leboyer M., Gillberg C., Bourgeron T. Abnormal melatonin synthesis in autism spectrum disorders. *Mol. Psychiatry*. 2008;13(1):90-98. DOI 10.1038/sj.mp.4002016

Modlinska K., Pisula W. The Norway rat, from an obnoxious pest to a laboratory pet. *eLife*. 2020;9:e50651. DOI 10.7554/eLife.50651

Moskaliuk V.S., Kozhemyakina R.V., Bazovkina D.V., Terenina E., Khomenko T.M., Volcho K.P., Salakhutdinov N.F., Kulikov A.V., Naumenko V.S., Kulikova E. On an association between fear-induced aggression and striatal-enriched protein tyrosine phosphatase (STEP) in the brain of Norway rats. *Biomed. Pharmacother*. 2022;147:112667. DOI 10.1016/j.biopha.2022.112667

Moskaliuk V.S., Kozhemyakina R.V., Khomenko T.M., Volcho K.P., Salakhutdinov N.F., Kulikov A.V., Naumenko V.S., Kulikova E.A. On associations between fear-induced aggression, *Bdnf* transcripts, and serotonin receptors in the brains of Norway rats: an influence of antiaggressive drug TC-2153. *Int. J. Mol. Sci*. 2023;24(2):983. DOI 10.3390/ijms24020983

Naumenko V.S., Kozhemjakina R.V., Plyusnina I.Z., Popova N.K. Expression of serotonin transporter gene and startle response in rats with genetically determined fear-induced aggression. *Bull. Exp. Biol. Med*. 2009;147(1):81-83. DOI 10.1007/s10517-009-0441-2

Oshchepkov D., Ponomarenko M., Klimova N., Chadaeva I., Bragin A., Sharypova E., Shikhevich S., Kozhemyakina R. A rat model of human behavior provides evidence of natural selection against underexpression of aggressiveness-related genes in humans. *Front. Genet*. 2019;10:1267. DOI 10.3389/fgene.2019.01267

Oshchepkov D., Chadaeva I., Kozhemyakina R., Zolotareva K., Khandaev B., Sharypova E., Ponomarenko P., Bogomolov A., Klimova N.V., Shikhevich S., Redina O., Kolosova N.G., Nazarenko M., Kolchanov N.A., Markel A., Ponomarenko M. Stress reactivity, susceptibility to hypertension, and differential expression of genes in hypertensive compared to normotensive patients. *Int. J. Mol. Sci*. 2022a;23(5):2835. DOI 10.3390/ijms23052835

Oshchepkov D., Chadaeva I., Kozhemyakina R., Shikhevich S., Sharypova E., Savinkova L., Klimova N.V., Tsukanov A., Levitsky V.G., Markel A.L. Transcription factors as important regulators of changes in behavior through domestication of gray rats: quantitative data from RNA sequencing. *Int. J. Mol. Sci*. 2022b;23(20):12269. DOI 10.3390/ijms232012269

Paxinos G., Watson C. The Rat Brain in Stereotaxic Coordinates. London: Acad. Press, Elsevier Inc., 2013.

Penning L.C., Vrieling H.E., Brinkhof B., Riemers F.M., Rothuizen J., Rutteman G.R., Hazewinkel H.A. A validation of 10 feline reference genes for gene expression measurements in snap-frozen tissues. *Vet. Immunol. Immunopathol*. 2007;120(3-4):212-222. DOI 10.1016/j.vetimm.2007.08.006

Perepechaeva M.L., Grishanova A.Y., Rudnitskaya E.A., Kolosova N.G. The mitochondria-targeted antioxidant SkQ1 downregulates aryl hydrocarbon receptor-dependent genes in the retina of OXYS rats with AMD-like retinopathy. *J. Ophthalmol*. 2014;2014:530943. DOI 10.1155/2014/530943

Plekanchuk V.S., Ryazanova M.A. Expression of glutamate receptor genes in the hippocampus and frontal cortex in GC rat strain with genetic catatonia. *J. Evol. Biochem. Phys*. 2021;57(1):156-163. DOI 10.1134/S0022093021010154

Plyusnina I., Oskina I. Behavioral and adrenocortical responses to open-field test in rats selected for reduced aggressiveness toward

humans. *Physiol. Behav*. 1997;61(3):381-385. DOI 10.1016/S0031-9384(96)00445-3

Popova N.K., Naumenko V.S., Plyusnina I.Z. Involvement of brain serotonin 5-HT1A receptors in genetic predisposition to aggressive behavior. *Neurosci. Behav. Physiol*. 2007;37(6):631-635. DOI 10.1007/s11055-007-0062-z

Popova N.K., Naumenko V.S., Kozhemyakina R.V., Plyusnina I.Z. Functional characteristics of serotonin 5-HT2A and 5-HT2C receptors in the brain and the expression of the 5-HT2A and 5-HT2C receptor genes in aggressive and non-aggressive rats. *Neurosci. Behav. Physiol*. 2010;40(4):357-361. DOI 10.1007/s11055-010-9264-x

Ryazanova M.A., Fedoseeva L.A., Ershov N.I., Efimov V.M., Markel A.L., Redina O.E. The gene-expression profile of renal medulla in ISIAH rats with inherited stress-induced arterial hypertension. *BMC Genet*. 2016;17(Suppl.3):151. DOI 10.1186/s12863-016-0462-6

Ryazanova M.A., Prokudina O.I., Plekanchuk V.S., Alekhina T.A. Expression of catecholaminergic genes in the midbrain and prepulse inhibition in rats with a genetic catatonia. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2017;21(7):798-803. DOI 10.18699/VJ17.296 (in Russian)

Ryazanova M.A., Plekanchuk V.S., Prokudina O.I., Makovka Y.V., Alekhina T.A., Redina O.E., Markel A.L. Animal models of hypertension (ISIAH rats), catatonia (GC rats), and audiogenic epilepsy (PM rats) developed by breeding. *Biomedicines*. 2023;11(7):1814. DOI 10.3390/biomedicines11071814

Sengupta P. The laboratory rat: relating its age with human's. *Int. J. Prev. Med*. 2013;4(6):624-630

Schmidt I. Metabolic diseases: the environment determines the odds, even for genes. *News Physiol. Sci*. 2002;17:115-121. DOI 10.1152/nips.01380.2001

Shikhevich S., Chadaeva I., Khandaev B., Kozhemyakina R., Zolotareva K., Kazachek A., Oshchepkov D., Bogomolov A., Klimova N.V., Ivanisenko V.A., Demenkov P., Mustafin Z., Markel A., Savinkova L., Kolchanov N.A., Kozlov V., Ponomarenko M. Differentially expressed genes and molecular susceptibility to human age-related diseases. *Int. J. Mol. Sci*. 2023;24(4):3996. DOI 10.3390/ijms24043996

Singh G., Bhat B., Jayadev M.S.K., Madhusudhan C., Singh A. mutTCPdb: a comprehensive database for genomic variants of a tropical country neglected disease-tropical calcific pancreatitis. *Database* (*Oxford*). 2018;2018:bay043. DOI 10.1093/database/bay043

Stefanova N.A., Kolosova N.G. The rat brain transcriptome: from infancy to aging and sporadic Alzheimer's disease-like pathology. *Int. J. Mol. Sci*. 2023;24(2):1462. DOI 10.3390/ijms24021462

Stefanova N.A., Maksimova K.Y., Rudnitskaya E.A., Muraleva N.A., Kolosova N.G. Association of cerebrovascular dysfunction with the development of Alzheimer's disease-like pathology in OXYS rats. *BMC Genomics*. 2018;19(Suppl.3):75. DOI 10.1186/s12864-018-4480-9

Stefanova N.A., Ershov N.I., Maksimova K.Y., Muraleva N.A., Tyumentsev M.A., Kolosova N.G. The rat prefrontal-cortex transcriptome: effects of aging and sporadic Alzheimer's disease-like pathology. *J. Gerontol. A Biol. Sci. Med. Sci*. 2019;74(1):33-43. DOI 10.1093/gerona/gly198

Stelzer G., Rosen N., Plaschkes I., Zimmerman S., Twik M., Fishilevich S., Stein T.I., Nudel R., Lieder I., Mazor Y., Kaplan S., Dahary D., Warshawsky D., Guan-Golan Y., Kohn A., Rappaport N., Safran M., Lancet D. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*. 2016;54:1.30.1-1.30.33. DOI 10.1002/cpbi.5

Stenson P.D., Mort M., Ball E.V., Shaw K., Phillips A., Cooper D.N. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet*. 2014;133(1):1-9. DOI 10.1007/s00439-013-1358-4

Sun S., Wang Y., Maslov A.Y., Dong X., Vijg J. SomaMutDB: a database of somatic mutations in normal human tissues. *Nucleic Acids Res*. 2022;50(D1):D1100-D1108. DOI 10.1093/nar/gkab914

Suzuki H., Han S.D., Lucas L.R. Increased 5-HT1B receptor density in the basolateral amygdala of passive observer rats exposed to aggression. *Brain Res. Bull*. 2010;83(1-2):38-43. DOI 10.1016/j.brainresbull.2010.06.007

Tain Y.L., Huang L.T., Chan J.Y., Lee C.T. Transcriptome analysis in rat kidneys: importance of genes involved in programmed hypertension. *Int. J. Mol. Sci*. 2015;16(3):4744-4758. DOI 10.3390/ijms16034744

Talarowska M., Szemraj J., Zajączkowska M., Galecki P. ASMT gene expression correlates with cognitive impairment in patients with recurrent depressive disorder. *Med. Sci. Monit*. 2014;20:905-912. DOI 10.12659/MSM.890160

Taylor J.R., Morshed S.A., Parveen S., Mercadante M.T., Scahill L., Peterson B.S., King R.A., Leckman J.F., Lombroso P.J. An animal model of Tourette's syndrome. *Am. J. Psychiatry*. 2002;159(4):657-660. DOI 10.1176/appi.ajp

Tharmalingam S., Khurana S., Murray A., Lamothe J., Tai T.C. Whole transcriptome analysis of adrenal glands from prenatal glucocorticoid programmed hypertensive rodents. *Sci. Rep*. 2020;10(1):18755. DOI 10.1038/s41598-020-75652-y

Trent S., Dean R., Veit B., Cassano T., Bedse G., Ojarikre O.A., Humby T., Davies W. Biological mechanisms associated with increased perseveration and hyperactivity in a genetic mouse model of neurodevelopmental disorder. *Psychoneuroendocrinology*. 2013;38(8):1370-1380. DOI 10.1016/j.psyneuen.2012.12.002

Wall V.L., Fischer E.K., Bland S.T. Isolation rearing attenuates social interaction-induced expression of immediate early gene protein products in the medial prefrontal cortex of male and female rats. *Physiol. Behav*. 2012;107(3):440-450. DOI 10.1016/j.physbeh.2012.09.002

Watanabe Y., Yoshida M., Yamanishi K., Yamamoto H., Okuzaki D., Nojima H., Yasunaga T., Okamura H., Matsunaga H., Yamanishi H. Genetic analysis of genes causing hypertension and stroke in spontaneously hypertensive rats: gene expression profiles in the kidneys. *Int. J. Mol. Med*. 2015;36(3):712-724. DOI 10.3892/ijmm.2015.2281

Wu H.M., Zhao C.C., Xie Q.M., Xu J., Fei G.H. TLR2-melatonin feedback loop regulates the activation of NLRP3 inflammasome in murine allergic airway inflammation. *Front. Immunol*. 2020;11:172. DOI 10.3389/fimmu.2020.00172

Xiao G., Wang T., Zhuang W., Ye C., Luo L., Wang H., Lian G., Xie L. RNA sequencing analysis of monocrotaline-induced PAH reveals dysregulated chemokine and neuroactive ligand receptor pathways. *Aging* (*Albany NY*). 2020;12(6):4953-4969. DOI 10.18632/aging.102922

Xie F., Wang L., Liu Y., Liu Z., Zhang Z., Pei J., Wu Z., Zhai M., Cao Y. ASMT regulates tumor metastasis through the circadian clock system in triple-negative breast cancer. *Front. Oncol*. 2020;10:537247. DOI 10.3389/fonc.2020.537247

Yang H., Zhang Z., Ding X., Jiang X., Tan L., Lin C., Xu L., Li G., Lu L., Qin Z., Feng X., Li M. RP58 knockdown contributes to hypoxia-ischemia-induced pineal dysfunction and circadian rhythm disruption in neonatal rats. *J. Pineal Res*. 2023;75(1):e12885. DOI 10.1111/jpi.12885

Ye J., Coulouris G., Zaretskaya I., Cutcutache I., Rozen S., Madden T.L. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 2012;13:134. DOI 10.1186/1471-2105-13-134

Yoshida M., Watanabe Y., Yamanishi K., Yamashita A., Yamamoto H., Okuzaki D., Shimada K., Nojima H., Yasunaga T., Okamura H., Matsunaga H., Yamanishi H. Analysis of genes causing hypertension and stroke in spontaneously hypertensive rats: gene expression profiles in the brain. *Int. J. Mol. Med*. 2014;33(4):887-896. DOI 10.3892/ijmm.2014.1631

Yuan X., Wu Q., Liu X., Zhang H., Xiu R. Transcriptomic profile analysis of brain microvascular pericytes in spontaneously hypertensive rats by RNA-Seq. *Am. J. Transl. Res*. 2018;10(8):2372-2386. PMID 30210677

Zhang H.F., Wang J.H., Wang Y.L., Gao C., Gu Y.T., Huang J., Wang J.H., Zhang Z. Salvianolic acid A protects the kidney against oxidative stress by activating the Akt/GSK-3β/Nrf2 signaling pathway and inhibiting the NF-κB signaling pathway in 5/6 nephrectomized rats. *Oxid. Med. Cell. Longev*. 2019;2019:2853534. DOI 10.1155/2019/2853534

Zhang Z., Silveyra E., Jin N., Ribelayga C.P. A congenic line of the C57BL/6J mouse strain that is proficient in melatonin synthesis. *J. Pineal Res*. 2018;65(3):e12509. DOI 10.1111/jpi.12509

ORCID ID

I.V. Chadaeva orcid.org/0000-0002-2724-5441
N.I. Ershov orcid.org/0000-0003-3423-3497
N.L. Podkolodnyy orcid.org/0000-0001-9132-7997
R. Kozhemyakina orcid.org/0000-0001-8948-1127
D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
E.Yu. Kondratyuk orcid.org/0000-0001-8672-7216
O.E. Redina orcid.org/0000-0003-0942-8460
O.S. Kozhevnikova orcid.org/0000-0001-6475-4061
N.A. Stefanova orcid.org/0000-0001-5127-5993
N.G. Kolosova orcid.org/0000-0003-2398-8544
A.L. Markel orcid.org/0000-0002-1550-1647
M.P. Ponomarenko orcid.org/0000-0003-1663-318X
D.Yu. Oshchepkov orcid.org/0000-0002-6097-5155

Acknowledgements. The authors are grateful to the Multi-Access Center "Bioinformatics" for access to computing resources under Project FWNR-2022-0020 and the Multi-Access Center "Conventional Animal Facility" for access to animals under Projects FWNR-2022-0019 and FWNR-2022-0015.

Conflict of interest. The authors declare no conflict of interest.

Received August 23, 2023. Revised September 11, 2023. Accepted September 15, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):807-814
DOI 10.18699/VJGB-23-93

Original Russian text https://vavilovj-icg.ru/

Application of the weighted histogram method for calculating the thermodynamic parameters of the formation of oligodeoxyribonucleotide duplexes

I.I. Yushin[1, 2], V.M. Golyshev[1, 2], D.V. Pyshnyi[1], A.A. Lomzov[1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
lomzov@niboch.nsc.ru

Abstract. To date, many derivatives and analogs of nucleic acids (NAs) have been developed. Some of them have found uses in scientific research and biomedical applications. Their effective use is based on the data about their properties. Some of the most important physicochemical properties of oligonucleotides are thermodynamic parameters of the formation of their duplexes with DNA and RNA. These parameters can be calculated only for a few NA derivatives: locked NAs, bridged oligonucleotides, and peptide NAs. Existing predictive approaches are based on an analysis of experimental data and the consequent construction of predictive models. The ongoing pilot studies aimed at devising methods for predicting the properties of NAs by computational modeling techniques are based only on knowledge about the structure of oligonucleotides. In this work, we studied the applicability of the weighted histogram analysis method (WHAM) in combination with umbrella sampling to the calculation of thermodynamic parameters of DNA duplex formation (changes in enthalpy ΔH°, entropy ΔS°, and Gibbs free energy ΔG°_{37}). A procedure was designed involving WHAM for calculating the hybridization properties of oligodeoxyribonucleotides. Optimal parameters for modeling and calculation of thermodynamic parameters were determined. The feasibility of calculation of ΔH°, ΔS°, and ΔG°_{37} was demonstrated using a representative sample of 21 oligonucleotides 4–16 nucleotides long with a GC content of 14–100 %. Error of the calculation of the thermodynamic parameters was 11.4, 12.9, and 11.8 % for ΔH°, ΔS°, and ΔG°_{37}, respectively, and the melting temperature was predicted with an average error of 5.5 °C. Such high accuracy of computations is comparable with the accuracy of the experimental approach and of other methods for calculating the energy of NA duplex formation. In this paper, the use of WHAM for computation of the energy of DNA duplex formation was systematically investigated for the first time. Our results show that a reliable calculation of the hybridization parameters of new NA derivatives is possible, including derivatives not yet synthesized. This work opens up new horizons for a rational design of constructs based on NAs for solving problems in biomedicine and biotechnology.
Key words: DNA; hybridization; thermodynamic parameters; Gibbs free energy; Weighted Histogram Analysis Method; WHAM; molecular dynamics.

For citation: Yushin I.I., Golyshev V.M., Pyshnyi D.V., Lomzov A.A. Application of the weighted histogram method for calculating the thermodynamic parameters of the formation of oligodeoxyribonucleotide duplexes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):807-814. DOI 10.18699/VJGB-23-93

Применение метода взвешенных гистограмм для расчета термодинамических параметров формирования комплексов олигодезоксирибонуклеотидов

И.И. Юшин[1, 2], В.М. Голышев[1, 2], Д.В. Пышный[1], А.А. Ломзов[1, 2]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
lomzov@niboch.nsc.ru

Аннотация. На сегодняшний день разработан широкий спектр производных и аналогов нуклеиновых кислот. Некоторые из них нашли применение при решении научно-исследовательских задач и задач биомедицины. Детальная информация о свойствах таких соединений является основой их эффективного использования. Одну из наиболее значимых физико-химических характеристик олигонуклеотидов – термодинамическую стабильность их дуплексов с ДНК и РНК – можно рассчитывать лишь для некоторых производных нуклеиновых кислот: LNA, мостиковых олигонуклеотидов и PNA. Существующие подходы основаны на анализе экспериментальных данных и построении прогностических моделей. Проводятся пилотные исследования, направленные на разработку методов прогнозирования свойств нуклеиновых кислот с использованием методов компьютерного моделирования, основанные только на знании структуры олигомеров. В данной работе исследована

применимость метода взвешенных гистограмм (WHAM) при анализе зонтичной выборки для расчета термодинамических параметров формирования ДНК-дуплексов: изменения энтальпии ΔH°, энтропии ΔS° и свободной энергии Гиббса ΔG°_{37}. Отработана процедура расчета гибридизационных свойств олигодезоксирибонуклеотидов с использованием метода взвешенных гистограмм. Подобраны оптимальные параметры проведения моделирования и расчета термодинамических параметров. На примере представительной выборки из 21 олигонуклеотида длиной от 4 до 16 нт и долей G/C пар от 14 до 100 % показана возможность расчета ΔH°, ΔS° и ΔG°_{37}. Ошибки расчета термодинамических параметров составляют 11.4, 12.9 и 11.8 % соответственно, а температура плавления прогнозируется со средней ошибкой 5.5 °C. Такая высокая точность расчетов сопоставима с экспериментальной и с другими прогностическими методами расчета энергии комплексообразования. В настоящей работе впервые систематически исследовано применение метода WHAM для расчета энергии формирования ДНК-дуплексов. Полученные результаты показывают потенциальную возможность достоверного расчета гибридизационных свойств новых, в том числе еще не синтезированных производных нуклеиновых кислот. Это открывает новые горизонты для рационального дизайна конструкций на основе нуклеиновых кислот для решения задач биомедицины и биотехнологии.
Ключевые слова: ДНК; гибридизация; термодинамические параметры; свободная энергия Гиббса; метод взвешенных гистограмм; WHAM; молекулярная динамика.

Introduction

To date, a wide range of derivatives and analogs of nucleic acids (NAs) have been developed, many of which have found applications in solving research problems and problems of biomedicine (e. g., (Wang et al., 2022)). Their effective use is possible due to the availability of detailed information about their physicochemical, molecular-biological, and biological properties. This information exists only for a limited number of derivatives of NAs such as locked nucleic acids (LNAs) (McTigue et al., 2004), peptide NAs (Griffin, Smith, 1998), phosphorothioate derivatives (Eckstein, 2014), phosphoramidate morpholino oligomers (Summerton, Weller, 1997), and bridged oligonucleotides (Lomzov et al., 2006). The development of approaches to the prediction of the properties of NAs, their analogs, and derivatives is absolutely necessary for rational design of oligonucleotide constructs in all the abovementioned applications. The availability of such approaches will greatly simplify both scientific research involving such compounds and the creation of commercial products, for example, molecular diagnostic systems or therapeutic NAs.

One of the key physicochemical properties of NA derivatives is their ability to form (and the efficiency of formation of) complexes with complementary sequences of DNA and RNA. Models have been devised to predictively calculate thermodynamic characteristics of the formation of duplex DNA structures (SantaLucia, Hicks, 2004), of duplex RNA structures (Xia et al., 1998), of hybrid DNA/RNA duplexes (Sugimoto et al., 1995; Banerjee et al., 2020), and of some NA derivatives: LNAs (McTigue et al., 2004), bridged oligonucleotides (Lomzov et al., 2006), and peptide NAs (Griffin, Smith, 1998). Such studies are based on analysis of experimental data about hybridization properties of these oligomers with consequent construction of predictive analytical models. In addition, pilot studies are being conducted that are aimed at designing techniques for reliable estimation of formation energy of NA complexes by computer modeling methods. The latter are promising from the standpoint of development of approaches to *a priori* prediction of properties of NA derivatives that have not yet been synthesized. In a recent paper, D. Dowerah and coworkers proposed a series of new analogs of LNAs with different linkers between O2′ and C4′ atoms (Dowerah et al., 2023). This work indicates high potential and demand for methods predicting the properties of modified NAs by means of only their chemical structure.

One well-established approach to the computation of Gibbs free energy is the weighted histogram analysis method (WHAM) combined with an analysis of umbrella sampling (e. g., (Kumar et al., 1992)). The general principle behind this calculation is to carry out molecular modeling by the umbrella sampling procedure and to analyze the resulting trajectories by the WHAM (Fig. 1). In molecular modeling by the umbrella sampling procedure, an additional (usually harmonic) potential is imposed on the system along the reaction coordinate (ξ), and this potential holds the system at position ξ_i ($i = 1 \ldots i_{max}$) with a certain force. For each umbrella sampling window (i), a histogram is obtained that represents a probability distribution along the reaction coordinate skewed by the holding potential. One of the most common techniques for calculating the potential of mean force (PMF) from histograms is the WHAM. Within this approach, researchers estimate statistical uncertainty of an unbiased (unshifted) probability distribution taking into account umbrella histograms and then compute the PMF that corresponds to the lowest uncertainty (Kumar et al., 1992). This approach allows to calculate free energy and other observable parameters (Grossfield, 2018).

In this work, we investigated the feasibility of calculating the formation energy of perfect DNA duplexes having various lengths and GC contents by the WHAM coupled with an analysis of an umbrella sample. The computation of the Gibbs free energy of duplex formation at different temperatures should enable us to calculate enthalpy (ΔH°) and entropy (ΔS°) contributions. By means of ΔH° and ΔS° values, it is possible to calculate the most illustrative and widely used characteristic for describing thermal stability of NA complexes: melting temperature (T_m).

Methods

The structure of DNA duplexes was created using the NAB program from the AmberTools18 software suite (Case et al., 2018). Starting structures had a B-form of the double helix.

A molecular dynamics (MD) simulation was performed in the AMBER18 software (Case et al., 2018) *via* parallel com-

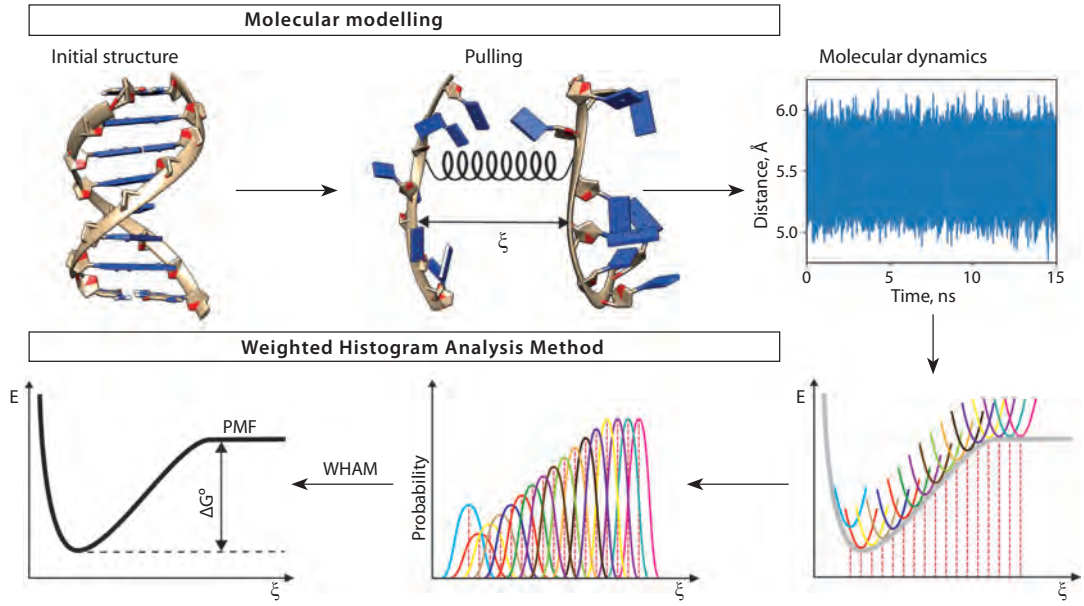


Fig. 1. The protocol for calculating the Gibbs free energy of formation of an NA double helix by the WHAM.

puting on central processing units and graphics accelerators. The ff99bsc0 force field was chosen to model DNA (Pérez et al., 2007). The MD simulation was carried out in an implicit water shell (Tsui, Case, 2000) at a fixed temperature in the range of 273 to 333 K with a step of 10 degrees by means of a Berendsen thermostat with a time constant of 10 ps (Omelyan, Kovalenko, 2013). To enable the step of integration of 2 fs motion equations, we employed the SHAKE algorithm.

The modeling procedure included eight stages:

1. Creating the structure of a DNA duplex and saving it in PDB format (with the help of the NAB program from the AmberTools18 software suite). Saving the structure in the amber file format (tleap).
2. Structure minimization for 10,000 steps (pmemd.cuda).
3. Stepwise heating of the system: from 0 to 100 K for 50 ps and from 100 K to a desired temperature (273 to 333 K in steps of 10 K) for 150 ps (pmemd.cuda). An integration time step of 0.5 fs was used.
4. Separation of two strands from 0 to 45 Å for 10 ns by applying 10 kcal/mol potential to the distance between the centers of mass of selected atoms of the strands (pmemd.cuda).
5. From the separation trajectory of the two DNA strands, extraction of structures for which the distance between the centers of mass was 0 to 45 (or 60) Å with a step of 0.5 Å (pmemd.cuda).
6. MD simulation of the extracted structures for 15 ns with the imposition of 10 kcal/mol harmonic potential on the distance between the centers of mass of the strands' selected atoms (pmemd.cuda).
7. Computation of interaction energy of the strands by the WHAM in the WHAM software (Grossfield, 2018). The number of points along the reaction coordinate for sampling of the free-energy profile was set to 150 (see below), and the convergence criterion of the WHAM was 10^{-6}.
8. Calculation of strand interaction energies *via* component-wise computation of free-energy changes based on MD simulation according to the generalized Born model (molecular mechanics/generalized Born surface area, MMGBSA) was performed using the MMPBSA.py module from the AmberTools18 software suite.

Molecular structures were visualized in the UCSF Chimera software (Pettersen et al., 2004).

Results and discussion

To refine the modeling protocol, a set of DNA oligomers having various lengths (4 to 16 bp) and GC contents (14 to 100 %) was chosen. Nucleotide sequences are given in the Table. The general protocol of the modeling and analysis is presented in Figure 1. We selected an approach where the distance between two DNA strands served as the reaction coordinate. That is, we carried out step-by-step separation of two strands in space and calculated the PMF depending on the distance between them. This approach combined with the WHAM makes it possible to determine the Gibbs energy of interaction between two strands directly in a computational experiment. If such an *in silico* experiment is conducted at different temperatures, it is possible to calculate a change in the enthalpy and entropy of complexation from a linear dependence of Gibbs free energy on temperature. At the first stage, it is necessary to determine the optimal parameters for performing such calculations.

As the reaction coordinate (ξ), we chose the distance (r) between centers of mass of C4′ atoms of all nucleotides from both strands. The initial distance was set to 0 Å in order to (i) examine the possibility of "compression" of the double helix, (ii) determine in the analysis the existence of a minimum of dependence $\Delta G^{\circ}_{T}(r)$, and (iii) compute the energy of complexation as the difference between the minimum and maximum of this dependence (see Fig. 1). Analyzing the separation of the strands' centers of mass showed that a maximum distance of 45 Å is sufficient for complete dissociation of the duplexes of oligonucleotides with a length of 4 to 9 bp, and for complexes

Duplex formation thermodynamic parameters calculated by the WHAM and the MMGBSA method and determined experimentally

Oligonucleotide sequence from the 5′ to the 3′end	WHAM					MMGBSA	Experimental values			
	ΔH°	ΔS°	ΔG°_{37}	T_m	R^2	ΔH°	ΔH°	ΔS°	ΔG°_{37}	T_m
AATTGGAC	–43.7	–115.6	–7.8	36.1	0.918	–77.4 ± 5.1	–56.9	–161	–6.9	31.7
ACGACCTC	–64.0	–169.1	–11.5	55.3	0.959	–85.1 ± 7.3	–59.8	–165	–8.6	40.5
AGAGCTCT	–64.3	–166.1	–12.8	62.5	0.958	–78.4 ± 8.1	–49.8	–134	–8.2	38.8
AGCATTAGACGGACCT	–166.0	–434.1	–31.3	87.8	0.960	–162.8 ± 7.9	–123.9	–335	–19.9	70.4
AGCCG	–39.0	–103.1	–7.0	29.8	0.923	–58.0 ± 5.8	–39.0	–108	–5.5	18.7
AGTTGC	–31.2	–82.0	–5.8	16.8	0.857	–65.4 ± 8.4	–37.0	–101	–5.7	19.0
ATATGGAC	–46.4	–130.6	–5.9	23.9	0.907	–77.6 ± 7.5	–53.8	–153	–6.5	28.0
CAAATAAAG	–67.9	–208.1	–3.4	17.5	0.963	–76.7 ± 8.6	–58.6	–168	–6.5	29.5
CACAG	–26.6	–71.2	–4.6	2.0	0.979	–56.6 ± 6.5	–33.7	–97	–3.6	1.7
CCGCGG	–60.8	–158.3	–11.7	57.2	0.932	–83.9 ± 7.6	–41.4	–106	–8.4	41.4
CGCG	–27.9	–68.3	–6.7	23.7	0.892		–36.3	–103	–4.5	9.1
CGCGCG	–45.5	–113.8	–10.2	52.9	0.905	–79.6 ± 6.7	–46.4	–121	–8.7	43.3
GCACCGAC	–92.3	–249.5	–14.9	62.2	0.986	–87.9 ± 7.6	–71.0	–196	–10.2	47.2
GCATGC	–58.8	–160.4	–9.1	43.1	0.948	–69.5 ± 7.9	–42.2	–117	–6.0	22.7
GCCCGGAC	–69.1	–182.8	–12.4	58.4	0.949	–94.9 ± 6.4	–61.4	–165	–10.3	48.9
GCCTGC	–48.3	–126.6	–9.0	44.0	0.915	–73.6 ± 8.6	–37.5	–100	–6.5	25.3
TACTGGAC	–62.7	–168.9	–10.3	48.9	0.934	–81.0 ± 7.7	–58.5	–165	–7.2	33.7
TCTATGCA	–44.3	–109.8	–10.3	54.4	0.813	–79.5 ± 6.5	–51.7	–145	–6.6	29.8
TGCGCA	–61.2	–162.3	–10.8	52.3	0.980	–76.9 ± 6.9	–42.5	–114	–7.3	31.2
TGTTGC	–41.1	–112.8	–6.1	23.6	0.979	–65.8 ± 7.9	–37.2	–101	–5.8	20.6
ACATTATTATTACA	–148.7	–442.7	–11.4	44.3	0.947	–125.4 ± 13.6	–89.9	–254	–11.1	48.3

Note. Sequence only one strand of each duplex is shown. Units of thermodynamic parameters ΔH° and ΔG°_{37}: kcal/mol, ΔS°: cal/(mol · K), and T_m: °C. R^2: Pearson's correlation coefficient for linear dependence ΔG°(*T*). Error of experimental ΔH°, ΔS°, ΔG°_{37} and T_m values is 10, 10, 8 %, and 0.5 °C, respectively.

of 14 and 16 bp in size, the distance should be increased to 60 Å. With the reaction coordinate chosen in this manner, the dissociation of the two strands for most oligonucleotide duplexes proceeded in accordance with the unzipping model (Cantor, Schimmel, 1980; Volkov, Solov'yov, 2009), which involves the unwinding of the double helix from one of the ends, or in accordance with a mixed shearing/unzipping mode (Mosayebi et al., 2015; Kurus, Dultsev, 2018). An example of alterations of oligonucleotide conformations along the reaction coordinate is given in Supplementary Material 1[1]. The mechanism of dissociation of duplexes in the current paper is not critically important because only two limiting states are being examined: a relaxed duplex structure and two noninteracting single-stranded oligonucleotides. The match between the helix–coil transition mechanism and the mechanisms observed by experimental methods confirmed the adequacy of the chosen approach for describing the dissociation of a DNA double helix.

According to generally accepted requirements for using the WHAM, it is necessary that the overlap between histograms be at least 20 %. Our analysis indicated that this is achieved at ~0.7 Å between adjacent simulation windows. An example of the dependence of distribution histograms on the distance between strands for duplex 5′-GCACCGAC-3′/5′-GTCGGTGC-3′ is given in Supplementary Material 2. We chose the reaction coordinate step of 0.5 Å to reliably meet this criterion.

When energy is calculated by the WHAM, an important parameter is the number of points ("bins") along the reaction coordinate that are chosen for sampling of the free energy profile. When the number of points was 100 or more (up to 1,000 partitions), a plateau was reached for the shape and position of the Gibbs free energy profiles (Fig. 2, *a* and Supplementary Material 3) and for the magnitude of the change in Gibbs free energy at different temperatures (see Fig. 2, *b*). At the same time, relative error values calculated by the bootstrap method (Grossfield, 2018) did not exceed 6 % (see Fig. 2, *c*).

Gibbs free energy at a certain temperature was calculated as the difference between a minimum and a maximum in the PMF profile: $\Delta G^\circ(T) = PMF_{min} - PMF_{max}$. To determine dependence of Gibbs free energy on temperature, the range from 273 to 333 K with a step of 10 K was chosen. The lower value was selected in accordance with the freezing temperature of water,

[1] Supplementary Materials 1–8 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Yushin_Engl_27_7.pdf

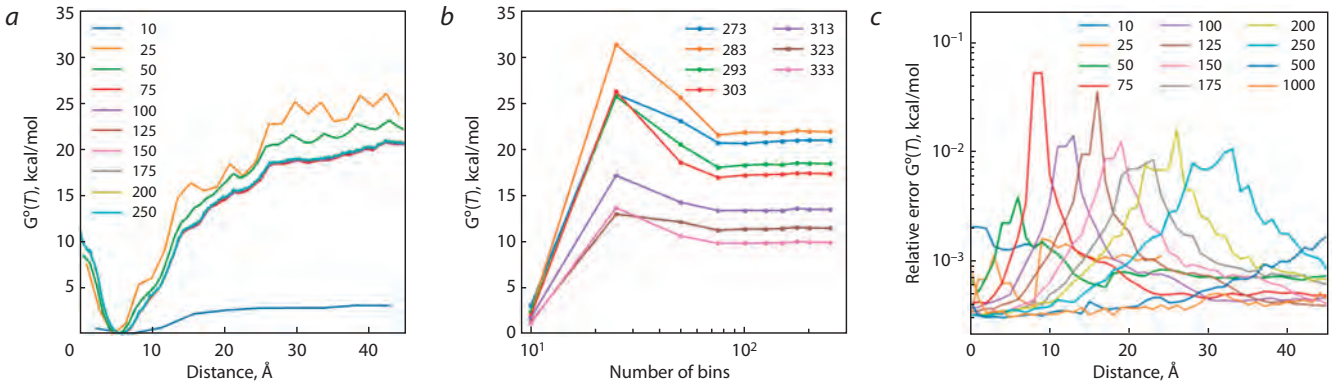


Fig. 2. Determination of parameters for the modeling and analysis of MD trajectories.
a, The dependence of the Gibbs free energy profile on the distance between the centers of mass of C4′ carbon atoms in two DNA strands for different numbers of points along the reaction coordinate that were chosen to sample the free-energy profile at 273 K. *b*, The dependence of the Gibbs energy of complexation on the number points along the reaction coordinate that were chosen to sample the free-energy profile. *c*, The dependence of relative error of Gibbs free energy computation on the distance between the centers of mass of C4′ carbon atoms from the two DNA strands for different numbers of points along the reaction coordinate that were chosen for sampling the free-energy profile.

and the upper value was chosen to limit the denaturation of an NA duplex in the modeling during the selected time range for short oligonucleotides. This range is wide enough for constructing dependence $\Delta G°(T)$ for reliable determination of $\Delta H°$ and $\Delta S°$ by linear regression analysis via the equation $\Delta G°(T) = \Delta H° - T\Delta S°$.

The trajectory length in the MD simulation at each fixed distance between the selected centers of mass and at a given temperature was set to 15 ns in order to obtain a minimally sufficient trajectory in an implicit water shell for the calculation of thermodynamic parameters (Lomzov et al., 2015). Thus, for each duplex, trajectories 15 ns long were obtained at 90 (or 120) different distances between the centers of mass at seven temperatures. Accordingly, trajectory length for each duplex ranged from 9.45 to 12.6 μs. The total length of trajectories for each complex was more than 200 μs.

By the WHAM, the dependence of the Gibbs energy of interaction of two oligonucleotides on the distance (r) between the centers of mass of C4′ atoms of each strand was calculated at seven tested temperatures for 21 studied duplexes (see the Table). A typical dependence of Gibbs energies of complex formation on r at temperatures of 273 to 333 K for the 5′-GCACCGAC-3′/5′-GTCGGTGC-3′ complex is depicted in Figure 3, *a*. The dependence of Gibbs free energy has a clear-cut minimum near 6 Å and increases with either an approach or dissociation of the double helix strands. During the dissociation, the dependence passes through a maximum and diminishes slightly. The maximum corresponds to the distance at which the interaction between the strands disappears.

To assess the adequacy of the modeling, we compared the geometry of the DNA double helix of the 5′-GCACCGAC-3′/5′-GTCGGTGC-3′ duplex in a relaxed form with literature data (Supplementary Material 4). All structural parameters are in good agreement with the data on Drew-Dickerson dodecamer (DDD, 5′-CGCGAATTCGCG-3′) structure determined experimentally by NMR spectroscopy (Protein Data Bank [PDB] ID: 1NAJ) and by X-ray crystallography (PDB ID: 1BNA).

Gibbs free energy of complexation was computed at various temperatures. It was established that $\Delta G°(T)$ is linear, with a high correlation coefficient R^2 of more than 0.83 and an average for all the analyzed complexes of 0.93 (see Fig. 3, *b* and the Table). Based on the obtained dependences (Supplementary Material 5), changes in the enthalpy and entropy of complexation were calculated next (see the Table). A comparison of thermodynamic parameters calculated by the WHAM with those determined experimentally (data from (Lomzov et al., 2015)) indicated a linear relation between them with high correlation coefficients R^2: 0.87, 0.82, 0.88, and 0.75 for $\Delta H°$, $\Delta S°$, $\Delta G°_{37}$ and T_m, respectively (Supplementary Material 6). As melting temperature, the temperature at which half of oligonucleotides are in a double-stranded state, and the remaining oligonucleotides are in a single-stranded state was chosen. T_m was computed from the thermodynamic parameters (Lomzov, Pyshnyi, 2012) as follows:

$$T_m = \Delta H°/(\Delta S° + R \ln\left(\frac{Ct}{4}\right)),$$

where R is the universal gas constant, and Ct is the total concentration of oligonucleotides in the system. Ct was set to 10 μM in accordance with typical experimental values.

The slope of the linear dependence of the thermodynamic parameters was found to be close to 0.5, and values of free terms of the linear dependence are substantial as compared to the analyzed values (see Supplementary Material 6). Therefore, as suggested in our previous papers (Lomzov et al., 2015; Golyshev et al., 2021), it is possible to apply linear corrections to calculated thermodynamic parameters $\Delta H°$ and $\Delta S°$. After this correction was applied, correlation coefficients for Gibbs free energy and melting temperatures improved considerably to 0.94 and 0.86, respectively (Fig. 4). In this context, the average absolute error of calculation of thermodynamic parameters became 11.4, 12.9, 11.8 %, and 5.5 °C for $\Delta H°$, $\Delta S°$, $\Delta G°_{37}$, and T_m, respectively. For our set of oligonucleotides, such error values for thermodynamic characteristics that have been obtained by the MMGBSA method in some studies (Lomzov et al., 2015; Golyshev et al., 2021) taking into account linear corrections are slightly lower: 7.6, 11.4, 10.6 %, and 4.3 °C. The accuracy of the computation of thermodynamic parameters in the present work is comparable to the accuracy of the experimental approach and to that of

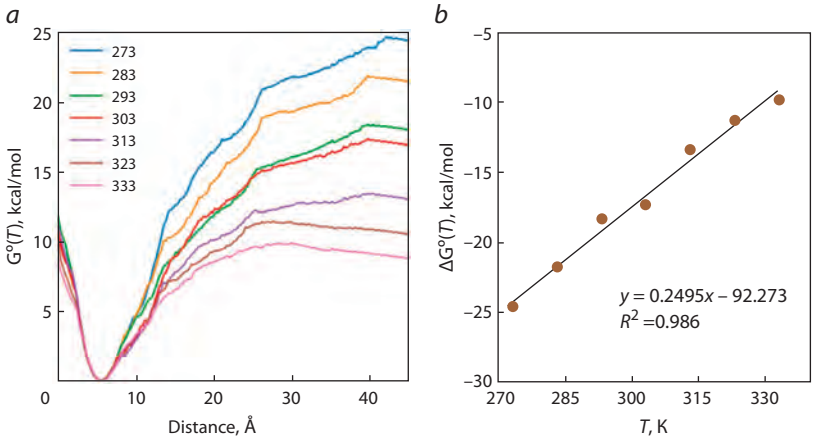


Fig. 3. The dependence of Gibbs free energy: *a*, on the distance between molecules at different temperatures (273, 283, 293, 303, 313, 323, and 333 K); *b*, on the temperature of model duplex 5′-GCACCGAC-3′/5′-GTCGGTGC-3′.

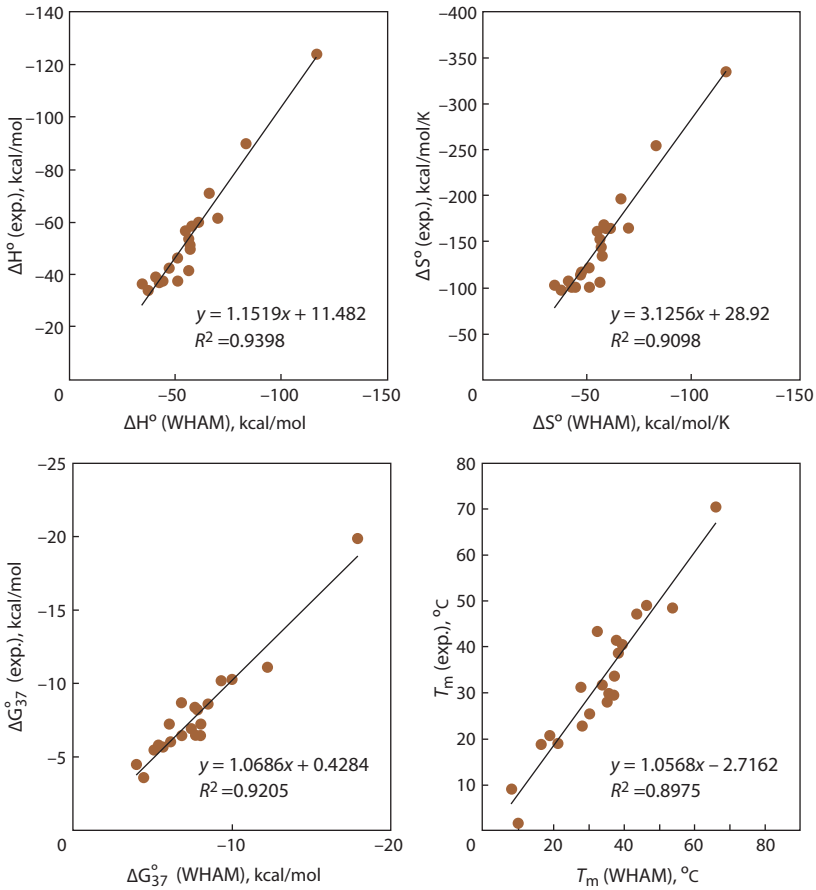


Fig. 4. Correlation of thermodynamic parameters ΔH°, ΔS°, ΔG°_{37}, and melting temperature of complexes – calculated by the WHAM taking into account linear corrections – with experimentally determined parameters (data from (Lomzov et al., 2015)).

the most common procedure for calculating the efficiency of oligonucleotides hybridization (the nearest-neighbor method): ~10 % for enthalpy and entropy and approximately 8 % for the Gibbs free energy of complexation (SantaLucia, Hicks, 2004; Lomzov et al., 2006).

To further check the quality of the results, the obtained trajectories were analyzed by the MMGBSA method, and the computed values were compared with the data of the WHAM and with values obtained by the MMGBSA method previously (Lomzov et al., 2015). The typical shape of the dependence of MMGBSA-calculated energy on the distance between the centers of mass of the strands' C4′ atoms proved to be similar to the dependence of Gibbs free energy on the distance depicted in Figure 3, *a* (Supplementary Material 7). At distances close to the maximum, the energy of formation of a DNA double helix reaches a plateau of zero, indicating the absence of interaction between the strands when this method of trajectory analysis is employed. The bottom of the potential well was observed in the region of 2–7 Å, which matches the relaxed form of the DNA double helix, and its global minimum near 7 Å is close to the minimum observed in Gibbs energy's dependence determined by the WHAM (see Fig. 3, *a* and Supplementary Material 7). There is a weak dependence of the complexation energy calculated by the MMGBSA method on temperature, implying a small change in heat capacity, ΔCp. It seems impossible to reliably determine the change in heat capacity by computational experiments owing to large values of calculation error.

The complexation enthalpy values computed in this work and those determined previously correlate well (R^2 = 0.97), with a slope close to unity (0.95) and the free term of the linear dependence close to zero (4 kcal/mol) (Supplementary Material 8, *a*). Additionally, a similar linear correlation was observed between the complexation enthalpy values calculated by the MMGBSA method and those determined by the WHAM in this work. Thus, the MD trajectories obtained in our paper are realistic.

One of the key aspects of the previously researched energy calculation by the MMGBSA method is the uncertainty associated with the structure of the single-stranded state of oligonucleotides. This state was extracted from the MD trajectory of a double helix. Nevertheless, this approach allowed to calculate the enthalpy of complexation with sufficient accuracy. In this work, during analysis by the WHAM, the single-stranded state of oligonucleotides was fairly well discernible in MD trajectories (as far as this can be done within the framework of the implicit water shell approximation and the force field in question). This approach yielded good results when the energy of double-helix formation was computed. Meanwhile, the main advantage of the WHAM is direct calculation of the change in the Gibbs free energy of DNA double-helix formation. This parameter

turned out to be linear in our calculations across a wide range of tested temperatures. This finding suggests that the modeling parameters selected by us and those included in the simulation and model analysis describe the physics of both double- and single-stranded DNA rather well. For the latter, this statement is supported by the conformation of oligonucleotides seen during the modeling of strands with a large distance between their centers of mass (see Supplementary Material 1). Oligonucleotides did not remain linear (in contrast to the duplex), and this observation was utilized in the MMGBSA analysis; they did not become completely disordered strands either but retained several heterocyclic bases in a row in stacking interactions. This finding is consistent with the persistent length of single-stranded regions of oligonucleotides, which is several nucleotides (depending on the GC content and on ionic strength of the solution) (e. g., (Chen et al., 2012)). Furthermore, the linear dependence $\Delta G°(T)$ observed in the procedure evaluated in the present paper makes it possible to directly calculate complexation entropy.

Nonetheless, the newly developed approach is far from perfect. In particular, for more accurate modeling of the structure and dynamics of DNA, it is necessary to employ the most modern force fields and an explicit water shell model. The analysis of force field parameters for such modeling is a separate, rather complicated task. Besides, the use of an explicit water shell greatly increases the complexity and duration of calculations. For instance, the main computational costs are incurred at the stage of MD calculations. For the 9 bp DNA duplex analyzed in detail in the current work, with the implicit water shell, the calculation speed for a modern video card (NVidia GTX 3080) is ~800 ns/day. Therefore, the computation time for one model duplex is 12 days. With an explicit water shell, the periodic cell being modeled will contain approximately 15–20 thousand molecules owing to the maximum distance between strands of 45 Å. This situation will reduce productivity to ~100 ns/day or approximately 3 months in total. In addition, in an explicit water shell, conformational mobility of DNA will significantly decrease, which will require extending trajectory length for each simulation window, thereby leading to higher computational costs. Nevertheless, such a complication seems necessary to improve the reliability/accuracy of the calculations.

Another promising avenue for the development of the proposed approach is its testing on known modified NAs as examples. This testing should answer the question of the applicability of our approach to the rational design of the chemical structure of new NA derivatives not yet chemically synthesized. The answer can help to solve specific problems of biomedicine and biotechnology. Our present analysis shows high potential and feasibility of the WHAM for calculating the formation energy of duplexes of NAs, their analogs, and derivatives.

Conclusion

A WHAM procedure for computing hybridization properties of oligodeoxyribonucleotides was refined here. Optimal parameters were selected for modeling and calculating thermodynamic parameters of the formation of DNA duplexes. By means of a representative sample of 21 oligonucleotides 4 to 16 nucleotides long with a GC content of 14 to 100 %, we demonstrated that calculating the enthalpy, entropy, and Gibbs free energy of the formation of oligonucleotide complexes by the WHAM is possible when MD trajectories are analyzed using the following reaction coordinate: the distance between the centers of mass of C4′ carbons of the two strands. A linear dependence of Gibbs free energy on the temperature at which the simulation is performed was documented. This finding enables researchers to compute the enthalpy and entropy of complexation *via* an analysis of WHAM results. The calculated thermodynamic parameters linearly correlate with experimentally determined values, with a high correlation coefficient R^2 (greater than 0.83). With a linear correction of this dependence, the error of calculation of thermodynamic parameters is comparable with the experimental one and amounts to 11.4, 12.9, and 11.8 % for $\Delta H°$, $\Delta S°$, and $\Delta G°_{37}$, while melting temperature is predicted with an average error of 5.5 °C. Thus, the use of the WHAM for calculating the formation energy of DNA duplexes was systematically investigated for the first time. High accuracy of such calculations was demonstrated, which is comparable with the accuracy of experimental and other techniques for computing the energy of complexation.

References

Banerjee D., Tateishi-Karimata H., Ohyama T., Ghosh S., Endoh T., Takahashi S., Sugimoto N. Improved nearest-neighbor parameters for the stability of RNA/DNA hybrids under a physiological condition. *Nucleic Acids Res.* 2020;48(21):12042-12054. DOI 10.1093/nar/gkaa572

Cantor C.R., Schimmel P.R. Biophysical Chemistry. Part I: The Conformation of Biological Macromolecules. New York: W.H. Freeman & Company, 1980

Case D.A., Walker R.C., Cheatham T.E., Simmerling C., Roitberg A., Merz K.M., Luo R., Darden T. Amber 18. Reference Manual. San Francisco: Univ. of California, 2018

Chen H., Meisburger S.P., Pabit S.A., Sutton J.L., Webb W.W., Pollack L. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. USA.* 2012;109(3):799-804. DOI 10.1073/pnas.1119057109

Dowerah D., Uppuladinne M.V.N., Sarma P.J., Biswakarma N., Sonavane U.B., Joshi R.R., Ray S.K., Namsa N.D., Deka R.C. Design of LNA analogues using a combined density functional theory and molecular dynamics approach for RNA therapeutics. *ACS Omega*. 2023;8(25):22382-22405. DOI 10.1021/acsomega.2c07860

Eckstein F. Phosphorothioates, essential components of therapeutic oligonucleotides. *Nucleic Acid Ther.* 2014;24(6):374-387. DOI 10.1089/nat.2014.0506

Golyshev V.M., Pyshnyi D.V., Lomzov A.A. Calculation of energy for RNA/RNA and DNA/RNA duplex formation by molecular dynamics simulation. *Mol. Biol.* 2021;55(6):927-940. DOI 10.1134/S002689332105006X

Griffin T.J., Smith L.M. An approach to predicting the stabilities of peptide nucleic acid:DNA duplexes. *Anal. Biochem.* 1998;260(1):56-63. DOI 10.1006/abio.1998.2686

Grossfield A. WHAM: the weighted histogram analysis method. 2018.

Kumar S., Rosenberg J.M., Bouzida D., Swendsen R.H., Kollman P.A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 1992;13(8):1011-1021. DOI 10.1002/jcc.540130812

Kurus N.N., Dultsev F.N. Determination of the thermodynamic parameters of DNA double helix unwinding with the help of mechanical methods. *ACS Omega*. 2018;3(3):2793-2797. DOI 10.1021/acsomega.7b01815

Lomzov A.A., Pyshnyi D.V. Considering the oligonucleotide secondary structures in thermodynamic and kinetic analysis of DNA duplex formation. *Biophysics* (*Oxf*). 2012;57(1):19-34. DOI 10.1134/S0006350912010137

Lomzov A.A., Pyshnaya I.A., Ivanova E.M., Pyshnyi D.V. Thermodynamic parameters for calculating the stability of complexes of bridged oligonucleotides. *Dokl. Biochem. Biophys.* 2006;409(1):211-215. DOI 10.1134/S1607672906040053

Lomzov A.A., Vorobjev Y.N., Pyshnyi D.V. Evaluation of the Gibbs free energy changes and melting temperatures of DNA/DNA duplexes using hybridization enthalpy calculated by molecular dynamics simulation. *J. Phys. Chem. B*. 2015;119(49):15221-15234. DOI 10.1021/acs.jpcb.5b09645

McTigue P.M., Peterson R.J., Kahn J.D. Sequence-dependent thermodynamic parameters for locked nucleic acid (LNA)-DNA duplex formation. *Biochemistry*. 2004;43(18):5388-5405. DOI 10.1021/bi035976d

Mosayebi M., Louis A.A., Doye J.P.K., Ouldridge T.E. Force-induced rupture of a DNA duplex: from fundamentals to force sensors. *ACS Nano*. 2015;9(12):11993-12003. DOI 10.1021/acsnano.5b04726

Omelyan I., Kovalenko A. Generalised canonical-isokinetic ensemble: speeding up multiscale molecular dynamics and coupling with 3D molecular theory of solvation. *Mol. Simul.* 2013;39(1):25-48. DOI 10.1080/08927022.2012.700486

Pérez A., Marchán I., Svozil D., Sponer J., Cheatham T.E., Laughton C.A., Orozco M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* 2007;92(11):3817-3829. DOI 10.1529/biophysj.106.097782

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* 2004;25(13):1605-1612. DOI 10.1002/jcc.20084

SantaLucia J., Hicks D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 2004;33(1):415-440. DOI 10.1146/annurev.biophys.32.110601.141800

Sugimoto N., Nakano S., Katoh M., Matsumura A., Nakamuta H., Ohmichi T., Yoneyama M., Sasaki M. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*. 1995;34(35):11211-11216. DOI 10.1021/bi00035a029

Summerton J., Weller D. Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev.* 1997;7(3):187-195. DOI 10.1089/oli.1.1997.7.187

Tsui V., Case D.A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers*. 2000;56(4):275-291. DOI 10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E

Volkov S.N., Solov'yov A.V. The mechanism of DNA mechanical unzipping. *Eur. Phys. J. D*. 2009;54(3):657-666. DOI 10.1140/epjd/e2009-00194-5

Wang F., Li P., Chu H.C., Lo P.K. Nucleic acids and their analogues for biomedical applications. *Biosensors*. 2022;12(2):93. DOI 10.3390/bios12020093

Xia T., SantaLucia J., Burkard M.E., Kierzek R., Schroeder S.J., Jiao X., Cox C., Turner D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*. 1998;37(42):14719-14735. DOI 10.1021/bi9809425

ORCID ID

I.I. Yushin orcid.org/0000-0001-5954-641X
V.M. Golyshev orcid.org/0000-0002-0521-6228
D.V. Pyshnyi orcid.org/0000-0002-2587-3719
A.A. Lomzov orcid.org/0000-0003-3889-9464

Acknowledgements. This work was supported by the Russian state-funded project for ICBFM SB RAS (grant number 121031300042-1).

Conflict of interest. The authors declare no conflict of interest.

Received July 15, 2023. Revised September 20, 2023. Accepted September 21, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):815-819
DOI 10.18699/VJGB-23-94

Original Russian text https://vavilovj-icg.ru/

Intratumor heterogeneity: models of malignancy emergence and evolution

R.A. Ivanov[1] ✉, S.A. Lashin[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ ivanovromanart@bionet.nsc.ru

Abstract. Cancer is a complex and heterogeneous disease characterized by the accumulation of genetic alterations that drive uncontrolled cell growth and proliferation. Evolutionary dynamics plays a crucial role in the emergence and development of tumors, shaping the heterogeneity and adaptability of cancer cells. From the perspective of evolutionary theory, tumors are complex ecosystems that evolve through a process of microevolution influenced by genetic mutations, epigenetic changes, tumor microenvironment factors, and therapy-induced changes. This dynamic nature of tumors poses significant challenges for effective cancer treatment, and understanding it is essential for developing effective and personalized therapies. By uncovering the mechanisms that determine tumor heterogeneity, researchers can identify key genetic and epigenetic changes that contribute to tumor progression and resistance to treatment. This knowledge enables the development of innovative strategies for targeting specific tumor clones, minimizing the risk of recurrence and improving patient outcomes. To investigate the evolutionary dynamics of cancer, researchers employ a wide range of experimental and computational approaches. Traditional experimental methods involve genomic profiling techniques such as next-generation sequencing and fluorescence *in situ* hybridization. These techniques enable the identification of somatic mutations, copy number alterations, and structural rearrangements within cancer genomes. Furthermore, single-cell sequencing methods have emerged as powerful tools for dissecting intratumoral heterogeneity and tracing clonal evolution. In parallel, computational models and algorithms have been developed to simulate and analyze cancer evolution. These models integrate data from multiple sources to predict tumor growth patterns, identify driver mutations, and infer evolutionary trajectories. In this paper, we set out to describe the current approaches to address this evolutionary complexity and theories of its occurrence.
Key words: cancer; evolution; heterogeneity.

For citation: Ivanov R.A., Lashin S.A. Intratumor heterogeneity: models of malignancy emergence and evolution. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):815-819. DOI 10.18699/VJGB-23-94

Внутриопухолевая гетерогенность: модели возникновения и эволюции злокачественных опухолей

Р.А. Иванов[1] ✉, С.А. Лашин[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ ivanovromanart@bionet.nsc.ru

Аннотация. Рак – сложное и гетерогенное заболевание, характеризующееся накоплением генетических изменений, которые приводят к неконтролируемому росту и пролиферации клеток. Эволюционная динамика играет решающую роль в возникновении и развитии раковых опухолей, формируя гетерогенность и адаптивность раковых клеток. С точки зрения теории эволюции опухоли представляют собой сложные экосистемы, которые развиваются в процессе микроэволюции под воздействием генетических мутаций, эпигенетических изменений и факторов микроокружения опухолей. Такая динамичная природа опухолей создает значительные проблемы для эффективного лечения рака, и ее понимание необходимо для разработки эффективных и персонализированных методов лечения. Раскрывая механизмы, определяющие гетерогенность опухоли, исследователи могут выявить ключевые генетические и эпигенетические изменения, которые способствуют прогрессированию опухоли и устойчивости к лечению. Эти знания позволяют разрабатывать инновационные стратегии воздействия на конкретные клоны опухоли, минимизируя риск рецидива и улучшая результаты лечения пациентов. Для изучения эволюционной динамики рака ученые используют широкий спектр экспериментальных и вычислительных подходов. Традиционные экспериментальные методы включают в себя геномное профилирование, такое как секвенирование нового поколения и флуоресцентная гибридизация *in situ*, и позволяют выявлять соматические мутации, изменения числа копий генов и структурные перестройки в геномах раковых опухолей. Помимо того, методы одноклеточного секвенирования стали мощным инструментом для изучения внутриопухолевой гетерогенности и отслеживания клональной эволюции. На основании экспериментальных данных разрабатываются

вычислительные модели и алгоритмы для моделирования и анализа эволюции рака. Эти модели объединяют данные из различных источников для предсказания закономерностей роста опухоли, выявления драйверных мутаций и построения эволюционных деревьев развития раковых клеток. В настоящей работе мы поставили задачу описать существующие на сегодняшний день подходы к изучению эволюционной динамики развития рака и теории ее возникновения.
Ключевые слова: злокачественные опухоли; эволюция; гетерогенность.

Evolutionary models of cancer

Cancer is a complex disease caused by the accumulation of genetic and epigenetic changes in normal cells, resulting in uncontrolled cell growth and tumor formation. Over the past few decades, it has become increasingly apparent that tumors are not static entities, but rather dynamic systems that undergo continuous evolution (Nowell, 1976; Merlo et al., 2006; Besse et al., 2018; Hausser, Alon, 2020; Vendramin et al., 2021). This evolutionary process shapes the heterogeneity and adaptability of cancer cells, posing significant challenges to effective cancer treatment. Tumor heterogeneity refers to the presence of different cell types in a tumor, commonly described as clones. In the context of oncology and evolutionary biomedicine, a clonal population is defined as a group of cancer cells that share a common origin and have similar genetic alterations. As these cells divide and accumulate additional mutations, they form separate clonal subpopulations in the tumor. This heterogeneity can manifest itself in various ways, such as differences in cell morphology (Meacham, Morrison, 2013; Robertson-Tessi et al., 2015; Haffner et al., 2021), differential gene expression of individual clones (Lüönd et al., 2021; Zhao et al., 2022), or their functional characteristics.

Clonal populations in cancer are commonly viewed as analogous to different species in the context of evolutionary biology (Vendramin et al., 2021). In the same way that different species evolve and adapt to their environment over time, clonal populations in a tumor evolve and adapt to their microenvironment. Genetic alterations emerging in these populations confer advantages or disadvantages in terms of growth, survival, and response to therapy, leading to selection and dominance of certain clones in the tumor.

Tumor heterogeneity represents a major treatment challenge because it can contribute to resistance to therapy, tumor recurrence after surgery, and the progression of metastasis (Morris et al., 2016). Currently, there are several theories regarding the mechanisms of the heterogeneity emergence in tumors.

The theory of clonal evolution is one of the earliest and most widely accepted theories that explains the occurrence of cancer heterogeneity. According to this theory, tumors originate from one or more transformed cells, the descendants of which acquire additional genetic mutations over time. These mutations lead to the formation of distinct clones with unique phenotypic characteristics. As the tumor grows, clones with advantageous traits are selected, resulting in the expansion and prevalence of these clones in the tumor population or their co-existence in the tumor depending on the type of cancer.

The concept of clonal evolution includes several models – linear, branching, and punctuated. In *the linear model*, mutations are acquired in a linear progression leading to more malignant stages of cancer (Fearon, Vogelstein, 1990). In the linear evolution model, new driver mutations provide such a strong selective advantage that they outcompete all previous clones due to the selective sweeping that occurs during tumor evolution. In *the branching evolution model*, clones diverge from a common ancestor and develop in parallel in a tumor tissue, giving rise to multiple clonal lineages (Gawad et al., 2014; Vosberg, Greif, 2019). In contrast to linear evolution, in the branching model of evolution, selective sweeps are rare, and multiple clonal populations evolve simultaneously because they all have increased adaptability. In this model, the magnitude of intratumor heterogeneity will fluctuate during tumor progression, but multiple clones are expected to be present at any given time of tumor sampling.

The neutral evolution model challenges the traditional view that all genetic alterations in cancer confer a selective advantage. According to this theory, most genetic mutations in cancer are neutral or nearly neutral, that is, they have no significant effect on tumor fitness (Williams et al., 2016; Furukawa, Kikuchi, 2020). Instead, the occurrence of heterogeneity is caused by random genetic drift, where neutral mutations randomly accumulate in different clones. Over time, these neutral mutations can become fixed within clones, leading to the observed intratumor heterogeneity.

It is worth noting that this theory is compatible with another popular theory of mutation accumulation – *punctuated evolution*, mentioned earlier in the text. According to this hypothesis, cancer cells are Goldschmidt's "hopeful monsters" (Graham, Sottoriva, 2017) – in which gradual and non-displayed changes in the genome lead to dramatic changes in the phenotype. Such a principle is evident in neoplasms in particular, since there are no obvious intermediate stages between healthy tissue and primary tumors. The intervals between the jumps, however, most likely represent the stages of neutral evolution. According to the same theory of punctuated evolution, the populations themselves may be in some kind of equilibrium with each other, maintaining several populations of clonal cancer cell lines in the tumor. After some time, one of the populations becomes a "hopeful monster" and in the case of a fitness-enhancing mutation, these clones occupy a larger part of the tumor, displacing the less adapted ones and increasing the size of the tumor itself.

Importantly, a number of studies have been reported that show that the development of an individual tumor does not necessarily follow a single pattern of clonal evolution and it can change during its development. Presumably, in the early stages of tumor development, it develops according to the linear evolution model, and once the tumor starts to actively grow, it switches to the branching model (Durrett et al., 2011; Vosberg, Greif, 2019). Moreover, several papers have shown that tumor evolution can follow both branching and punctuated models simultaneously – when clones with gene copy number changes follow the punctuated model and clones with

point mutations follow the branching model (Baca et al., 2013; Wang et al., 2014).

Another common theory on the origin of heterogeneity is the **cancer stem cell theory**, which suggests that tumors are hierarchically organized structures and only a small population of cancer stem cells (CSCs) determines tumor growth and heterogeneity (Reya et al., 2001; Lee et al., 2022). CSCs have the ability to self-renew and differentiate, similar to normal stem cells. These cells are capable of generating both other CSCs and non-CSC progeny, which in theory contributes to the cellular diversity seen in tumors. An important aspect of this theory is the hierarchy of cancer cells – normal cancer cells are incapable of differentiation and somatic mutations in them have a less significant clinical effect due to a lower ability to reproduce, while the main pathological significance is due to CSCs with different degrees of pluripotency. The occurrence of heterogeneity in this model is explained by asymmetric division of CSCs, which can lead to the appearance of different CSC clones with different phenotypic properties. It is worth noting that so far CSCs have only been found in a limited number of tumor types, particularly in hematologic tumors (Bonnet, Dick, 1997; Zarzynska, 2017; Hata et al., 2018; Lee et al., 2022), but in these instances they may be a major factor in malignant tumor recurrence after treatment (Walcher et al., 2020).

The theory of microenvironmental selection suggests that the tumor microenvironment plays an important role in shaping tumor heterogeneity. The interaction between cancer cells and the surrounding microenvironment, which includes immune cells, stromal cells, and extracellular matrix components, may exert selective pressure on tumor cells (Augustin et al., 2020). Microenvironmental factors such as hypoxia, inflammation, and nutrient availability can influence tumor growth, angiogenesis, and metastasis (Mumenthaler et al., 2015; Roma-Rodrigues et al., 2019). This selective pressure favors the survival and reproduction of specific clones with advantageous traits that allow them to adapt to the microenvironment.

Among the factors of the microenvironment, the immune system plays a particularly important role. The action of immune cells has a double function in cancer development: it can both inhibit tumor growth and promote tumor progression. Immune checkpoint mechanisms recognize and destroy cancer cells, preventing tumor formation. However, tumors can evade the immune response through a variety of mechanisms, leading to the immune response acting as a natural selection factor for clonal populations and thus selecting the most resistant clonal populations with altered antigens, which directly affects the severity of the disease and the efficacy of immunotherapy.

Finally, **the theory of epigenetic plasticity** suggests that, in addition to genetic abnormalities, epigenetic alterations also play a significant role in causing tumor heterogeneity (Flavahan et al., 2017; Yao et al., 2020). Epigenetic modifications, such as DNA methylation and histone modifications, can dynamically regulate gene expression patterns and cellular phenotypes. According to this theory, cancer cells possess an epigenetic landscape plasticity that allows for reversible and dynamic changes in gene expression. These epigenetic changes can give rise to different clones with distinct phenotypic characteristics, contributing to intratumor heterogeneity.

Approaches to the study of evolutionary characteristics in heterogeneous tumors

To study the evolutionary features of heterogeneous tumors, it is imperative for the researcher to be able to qualitatively and quantitatively assess different clonal populations. In the next section, we present a number of analysis methods that are currently used to study tumor heterogeneity.

The population genetics approach is one way to theoretically study heterogeneous tumor communities. According to population genetics, the evolution of a population relies on two factors: the mutation rate and the effective population size. The mutation rate refers to the expected number of genetic mutations per individual replication event and directly impacts the diversity within a population. The effective population size determines the population's capacity to maintain this diversity. In tumors, the effective size is defined as the total number of cancer cells, but it is also possible to exclude some groups of cancer cells from this number – if, for example, a CSC-induced tumor is modeled, which would be the main cause of tumor growth. Of course, such an approach requires the use of single-cell sequencing of tumors. Due to the complexity and high cost of this method, classical population genetics analysis has only been performed in a few papers so far (Navin, 2015; Losic et al., 2020; Heinrich et al., 2021; Deng et al., 2023).

Since single-cell sequencing methods have only recently become available, much of the work has focused on studying heterogeneity using bulk next-generation sequencing methods on tumor samples. This approach has an obvious problem: it is difficult to directly identify the clonal architecture of a tumor in the data obtained from such samples. Therefore, using this approach, researchers have to make certain assumptions and modifications to experimental methods. One of them is to increase the sequencing depth to estimate the frequencies of mutant alleles (Koh et al., 2021). To analyze tumor populations, statistical methods are used to normalize these frequencies and cluster genotypes to identify identical clonal populations. Diversity characterizations like the Shannon diversity index and Simpson index are often employed in such studies. However, a drawback of this approach is its inability to distinguish between populations if they have similar mutant allele frequencies.

Another modification is multiregional sequencing, in which samples are collected from multiple tumor sites. In particular, this method allows us to assess the difference in heterogeneity in patients with multiple metastatic tumors, which in the context of diversity can be perceived as a population of clones with prolonged physical isolation.

The most promising techniques for experimental assessment of heterogeneity are methods of single cell analysis, as they allow us to judge the individual differences of clones at the genetic and phenotypic levels. Immunofluorescence *in situ* hybridization (iFISH) is one such technique. Through the use of fluorescently labeled DNA probes that hybridize with complementary target sequences, FISH allows the detection of genetic alterations, chromosomal rearrangements and gene amplifications with high specificity and sensitivity. *In situ* FISH (iFISH) is the implementation of FISH directly on tissue sections while preserving the spatial organization of cells in the tumor microenvironment (Gertz et al., 2016). However,

the iFISH method is low-throughput and does not allow for the investigation of heterogeneity at the full-genome level.

In contrast to the method described above, single-cell sequencing (scDNA-seq and scRNA-seq) allows us to determine the pattern of genetic diversity, gene expression in each individual cell and decipher its intercellular signaling networks. These methods provide a clear picture not only of the mechanisms of intratumor heterogeneity, but also of intercellular interactions through ligand-receptor signaling.

Conclusion

Understanding the evolution and heterogeneity of malignant tumors is crucial for improving cancer diagnosis and developing treatment strategies. Many molecular genetic techniques, with their advantages and disadvantages, have been developed to study the genetic and phenotypic characteristics of cancer clone populations. Next-generation sequencing can provide a comprehensive view of the genomic landscape of a tumor, but there is a risk of missing rare clones. Single-cell sequencing can identify rare clones and reconstruct clonal lineages, but is technically challenging and expensive. Methods such as iFISH provide spatial information but have limited target coverage and are low throughput.

Based on the data obtained using such methods, various models have been proposed to explain the dynamic nature of tumor evolution, including models of clonal evolution, cancer stem cells, models of microenvironmental impact, and epigenetic factors. Each of them provides valuable insights into the mechanisms behind tumor heterogeneity and the emergence of drug resistance.

Moreover, the development of mathematical and computational models of clonal evolution and algorithms for analyzing large-scale genomic data could enhance the ability to interpret and extract meaningful information from complex datasets of malignancies. These tools would potentially allow researchers to identify key driver events, track evolutionary dynamics, and more accurately predict the effects of treatment.

References

Augustin R.C., Delgoffe G.M., Najjar Y.G. Characteristics of the tumor microenvironment that influence immune cell functions: hypoxia, oxidative stress, metabolic alterations. *Cancers* (*Basel*). 2020; 12(12):3802. DOI 10.3390/cancers12123802

Baca S.C., Prandi D., Lawrence M.S., Mosquera J.M., Romanel A., Drier Y., Park K., Kitabayashi N., MacDonald T.Y., Ghandi M., Van Allen E., Kryukov G.V., Sboner A., Theurillat J.-P., Soong T.D., Nickerson E., Auclair D., Tewari A., Beltran H., Onofrio R.C., Boysen G., Guiducci C., Barbieri C.E., Cibulskis K., Sivachenko A., Carter S.L., Saksena G., Voet D., Ramos A.H., Winckler W., Cipicchio M., Ardlie K., Kantoff P.W., Berger M.F., Gabriel S.B., Golub T.R., Meyerson M., Lander E.S., Elemento O., Getz G., Demichelis F., Rubin M.A., Garraway L.A. Punctuated evolution of prostate cancer genomes. *Cell*. 2013;153(3):666-677. DOI 10.1016/j.cell.2013.03.021

Besse A., Clapp G.D., Bernard S., Nicolini F.E., Levy D., Lepoutre T. Stability analysis of a model of interaction between the immune system and cancer cells in chronic myelogenous leukemia. *Bull. Math. Biol.* 2018;80(5):1084-1110. DOI 10.1007/s11538-017-0272-7

Bonnet D., Dick J.E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* 1997;3(7):730-737. DOI 10.1038/nm0797-730

Deng G., Zhang X., Chen Y., Liang S., Liu S., Yu Z., Lü M. Single-cell transcriptome sequencing reveals heterogeneity of gastric cancer: progress and prospects. *Front. Oncol.* 2023;13:1074268. DOI 10.3389/fonc.2023.1074268

Durrett R., Foo J., Leder K., Mayberry J., Michor F. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*. 2011;188(2):461-477. DOI 10.1534/genetics.110.125724

Fearon E.R., Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell*. 1990;61(5):759-767. DOI 10.1016/0092-8674(90)90186-I

Flavahan W.A., Gaskell E., Bernstein B.E. Epigenetic plasticity and the hallmarks of cancer. *Science*. 2017;357(6348):eaal2380. DOI 10.1126/science.aal2380

Furukawa Y., Kikuchi J. Molecular basis of clonal evolution in multiple myeloma. *Int. J. Hematol.* 2020;111(4):496-511. DOI 10.1007/s12185-020-02829-6

Gawad C., Koh W., Quake S.R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA*. 2014;111(50):17947-17952. DOI 10.1073/pnas.1420822111

Gertz E.M., Chowdhury S.A., Lee W.-J., Wangsa D., Heselmeyer-Haddad K., Ried T., Schwartz R., Schäffer A.A. FISHtrees 3.0: tumor phylogenetics using a ploidy probe. *PLoS One*. 2016;11(6): e0158569. DOI 10.1371/journal.pone.0158569

Graham T.A., Sottoriva A. Measuring cancer evolution from the genome. *J. Pathol.* 2017;241(2):183-191. DOI 10.1002/path.4821

Haffner M.C., Zwart W., Roudier M.P., True L.D., Nelson W.G., Epstein J.I., De Marzo A.M., Nelson P.S., Yegnasubramanian S. Genomic and phenotypic heterogeneity in prostate cancer. *Nat. Rev. Urol.* 2021;18(2):79-92. DOI 10.1038/s41585-020-00400-w

Hata M., Hayakawa Y., Koike K. Gastric stem cell and cellular origin of cancer. *Biomedicines*. 2018;6(4):100. DOI 10.3390/biomedicines6040100

Hausser J., Alon U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nat. Rev. Cancer*. 2020;20(4):247-257. DOI 10.1038/s41568-020-0241-6

Heinrich S., Craig A.J., Ma L., Heinrich B., Greten T.F., Wang X.W. Understanding tumour cell heterogeneity and its implication for immunotherapy in liver cancer using single-cell analysis. *J. Hepatol.* 2021;74(3):700-715. DOI 10.1016/j.jhep.2020.11.036

Koh G., Degasperi A., Zou X., Momen S., Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer*. 2021;21(10):619-637. DOI 10.1038/s41568-021-00377-7

Lee T.K.-W., Guan X.-Y., Ma S. Cancer stem cells in hepatocellular carcinoma – from origin to clinical implications. *Nat. Rev. Gastroenterol. Hepatol.* 2022;19(1):26-44. DOI 10.1038/s41575-021-00508-3

Losic B., Craig A.J., Villacorta-Martin C., Martins-Filho S.N., Akers N., Chen X., Ahsen M.E., von Felden J., Labgaa I., D'Avola D., Allette K., Lira S.A., Furtado G.C., Garcia-Lezana T., Restrepo P., Stueck A., Ward S.C., Fiel M.I., Hiotis S.P., Gunasekaran G., Sia D., Schadt E.E., Sebra R., Schwartz M., Llovet J.M., Thung S., Stolovitzky G., Villanueva A. Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Commun.* 2020;11(1):291. DOI 10.1038/s41467-019-14050-z

Lüönd F., Tiede S., Christofori G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br. J. Cancer*. 2021;125(2):164-175. DOI 10.1038/s41416-021-01328-7

Meacham C.E., Morrison S.J. Tumour heterogeneity and cancer cell plasticity. *Nature*. 2013;501(7467):328-337. DOI 10.1038/nature12624

Merlo L.M.F., Pepper J.W., Reid B.J., Maley C.C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*. 2006;6(12):924-935. DOI 10.1038/nrc2013

Morris L.G.T., Riaz N., Desrichard A., Şenbabaoğlu Y., Hakimi A.A., Makarov V., Reis-Filho J.S., Chan T.A. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*. 2016;7(9):10051-10063. DOI 10.18632/oncotarget.7067

Mumenthaler S.M., Foo J., Choi N.C., Heise N., Leder K., Agus D.B., Pao W., Michor F., Mallick P. The impact of microenvironmental

heterogeneity on the evolution of drug resistance in cancer cells. *Cancer Inform.* 2015;14(Suppl.4):19-31. DOI 10.4137/CIN.S19338

Navin N.E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* 2015;25(10):1499-1507. DOI 10.1101/gr.191098.115

Nowell P. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23-28. DOI 10.1126/science.959840

Reya T., Morrison S.J., Clarke M.F., Weissman I.L. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. DOI 10.1038/35102167

Robertson-Tessi M., Gillies R.J., Gatenby R.A., Anderson A.R.A. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer Res.* 2015;75(8):1567-1579. DOI 10.1158/0008-5472.CAN-14-1428

Roma-Rodrigues C., Mendes R., Baptista P., Fernandes A. Targeting tumor microenvironment for cancer therapy. *Int. J. Mol. Sci.* 2019;20(4):840. DOI 10.3390/ijms20040840

Vendramin R., Litchfield K., Swanton C. Cancer evolution: Darwin and beyond. *EMBO J*. 2021;40(18):e108389. DOI 10.15252/embj.2021108389

Vosberg S., Greif P.A. Clonal evolution of acute myeloid leukemia from diagnosis to relapse. *Genes Chromosomes Cancer*. 2019;58(12):839-849. DOI 10.1002/gcc.22806

Walcher L., Kistenmacher A.-K., Suo H., Kitte R., Dluczek S., Straub A., Blaudszun A.-R., Yevsa T., Fricke S., Kossatz-Boehlert U. Cancer stem cells-origins and biomarkers: perspectives for targeted personalized therapies. *Front. Immunol.* 2020;11:1280. DOI 10.3389/fimmu.2020.01280

Wang Y., Waters J., Leung M.L., Unruh A., Roh W., Shi X., Chen K., Scheet P., Vattathil S., Liang H., Multani A., Zhang H., Zhao R., Michor F., Meric-Bernstam F., Navin N.E. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*. 2014;512(7513):155-160. DOI 10.1038/nature13600

Williams M.J., Werner B., Barnes C.P., Graham T.A., Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* 2016;48(3):238-244. DOI 10.1038/ng.3489

Yao J., Chen J., Li L.-Y., Wu M. Epigenetic plasticity of enhancers in cancer. *Transcription*. 2020;11(1):26-36. DOI 10.1080/21541264.2020.1713682

Zarzynska J.M. The role of stem cells in breast cancer. In: Breast Cancer – From Biology to Medicine. InTech, 2017. DOI 10.5772/66904

Zhao T., Chiang Z.D., Morriss J.W., LaFave L.M., Murray E.M., Del Priore I., Meli K., Lareau C.A., Nadaf N.M., Li J., Earl A.S., Macosko E.Z., Jacks T., Buenrostro J.D., Chen F. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature*. 2022;601(7891):85-91. DOI 10.1038/s41586-021-04217-4

ORCID ID

R.A. Ivanov orcid.org/0000-0002-4369-356X
S.A. Lashin orcid.org/0000-0003-3138-381X

Acknowledgements. The work was supported by Budget Project No. FWNR-2022-0020.

Conflict of interest. The authors declare no conflict of interest.

Received July 13, 2023. Revised August 7, 2023. Accepted August 17, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):820-828
DOI 10.18699/VJGB-23-95

Original Russian text https://vavilovj-icg.ru/

Search for differentially methylated regions in ancient and modern genomes

D.D. Borodko ✉, S.V. Zhenilo, F.S. Sharko

Federal Research Center "Fundamentals of Biotechnology" of the Russian Academy of Sciences, Moscow, Russia
✉ daria.borodko@gmail.com

Abstract. Currently, active research is focused on investigating the mechanisms that regulate the development of various pathologies and their evolutionary dynamics. Epigenetic mechanisms, such as DNA methylation, play a significant role in evolutionary processes, as their changes have a faster impact on the phenotype compared to mutagenesis. In this study, we attempted to develop an algorithm for identifying differentially methylated regions associated with metabolic syndrome, which have undergone methylation changes in humans during the transition from a hunter-gatherer to a sedentary lifestyle. The application of existing whole-genome bisulfite sequencing methods is limited for ancient samples due to their low quality and fragmentation, and the approach to obtaining DNA methylation profiles differs significantly between ancient hunter-gatherer samples and modern tissues. In this study, we validated DamMet, an algorithm for reconstructing ancient methylomes. Application of DamMet to Neanderthal and Denisovan genomes showed a moderate level of correlation with previously published methylation profiles and demonstrated an underestimation of methylation levels in the reconstructed profiles by an average of 15–20 %. Additionally, we developed a new Python-based algorithm that allows for the comparison of methylomes in ancient and modern samples, despite the absence of methylation profiles in modern bone tissue within the context of obesity. This analysis involves a two-step data processing approach, where the first step involves the identification and filtration of tissue-specific methylation regions, and the second step focuses on the direct search for differentially methylated regions in specific areas associated with the researcher's target condition. By applying this algorithm to test data, we identified 38 differentially methylated regions associated with obesity, the majority of which were located in promoter regions. The pipeline demonstrated sufficient efficiency in detecting these regions. These results confirm the feasibility of reconstructing DNA methylation profiles in ancient samples and comparing them with modern methylomes. Furthermore, possibilities for further methodological development and the implementation of a new step for studying differentially methylated positions associated with evolutionary processes are discussed.
Key words: ancient DNA; methylation; epigenetics; DamMet; DMR.

For citation: Borodko D.D., Zhenilo S.V., Sharko F.S. Search for differentially methylated regions in ancient and modern genomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):820-828. DOI 10.18699/VJGB-23-95

Поиск дифференциально метилированных регионов в геномах древних и современных людей

Д.Д. Бородко ✉, С.В. Женило, Ф.С. Шарко

Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Россия
✉ daria.borodko@gmail.com

Аннотация. В настоящее время активно исследуются механизмы, регулирующие развитие различных патологий и их эволюционную динамику. Эпигенетические механизмы, такие как метилирование, играют значимую роль в эволюционных процессах, поскольку их изменения гораздо быстрее отражаются на фенотипе, чем результаты мутагенеза. В данном исследовании мы предприняли попытку разработать алгоритм для выявления дифференциально метилированных областей, связанных с метаболическим синдромом, которые изменили свое метилирование у человека при переходе от охоты и собирательства к оседлой жизни. Применение существующих методов полногеномного бисульфитного секвенирования ограничено для древних образцов из-за их низкого качества и фрагментации, и подход к получению профилей метилирования охотников-собирателей значительно отличается от подходов, используемых для современных тканей. В этой работе мы валидировали DamMet – алгоритм, реконструирующий древние метиломы. Применение DamMet к геномам неандертальца и денисовца показало средний уровень корреляции с профилями метилирования, опубликованными ранее, а также продемонстрировало занижение уровня метилирования реконструированных профилей в среднем на 15–20 %. Также мы разработали новый алгоритм на языке Python, позволяющий сравнивать метиломы в древних и современных образцах, несмотря на отсутствие профилей метилирования современных образцов костной ткани в контексте ожирения. Такой анализ подразумевает двухступенчатую обработку данных, где

на первом этапе происходит идентификация тканеспецифичных областей метилирования и их фильтрация, а на втором этапе осуществляется непосредственно поиск дифференциально метилированных регионов в заданных областях, ассоциированных с интересующим исследователя заболеванием. В результате использования алгоритма на тестовых данных мы обнаружили 38 дифференциально метилированных регионов, ассоциированных с ожирением, большая часть которых принадлежала промоторным областям, и разработанный пайплайн показал достаточную эффективность в их поиске. Эти результаты подтверждают возможность восстановления профилей метилирования в древних образцах и их сравнения с современными метиломами. Также обсуждаются возможности дальнейшего развития методологии и внедрения нового шага, позволяющего изучать дифференциально метилированные позиции, связанные с эволюционными процессами.
Ключевые слова: древняя ДНК; метилирование; эпигенетика; DamMet; ДМР.

Introduction

Lately, increasing attention is being paid to the study of mechanisms regulating the development of various pathologies and their evolutionary dynamics (Briggs et al., 2009a; Niiranen et al., 2022). Epigenetic mechanisms, such as methylation, play a particularly important role in this process since they are capable of inducing phenotypic changes much faster than conventional mutagenesis processes (Jablonka, Raz, 2009; Feinberg, Irizarry, 2010; Zhur et al., 2021). The main goal of this study was to identify differentially methylated regions (DMRs) associated with metabolic syndrome, which could potentially serve as targets for epigenetic therapy of metabolic syndrome.

Nowadays, scientists are often hindered from conducting evolutionary research due to the lack of suitable methods for comparing DNA profiles of ancient and modern samples. Laboratory protocols used to obtain these profiles significantly differ from one another, each having its peculiarities and errors. Ancient DNA (aDNA) is often found in a fragmented state, and over time, natural molecule degradation and spontaneous deamination of nitrogenous bases occur, limiting the availability of high-quality data (Briggs et al., 2007, 2009b). To address this issue, a specific sample processing protocol was developed, which uses uracil-DNA glycosylase (UDG) and endonuclease combination (known as USER-treatment) to facilitate the extraction of methylation profiles and enhance their distinguishability (Briggs et al., 2010). Additionally, several programs have been developed that allow the calculation of methylation levels in ancient samples, the sequences of which were sequenced using the USER treatment (Gokhman et al., 2014; Orlando et al., 2015; Hanghøj et al., 2019).

At present, two methylation reconstruction algorithms tailored for ancient samples are available, characterized by their command-line functionality and user-friendliness. The antecedent algorithm, epiPALEOMIX, draws its foundation from the initial historical approach to methylation reconstruction, as first elucidated by D. Gokhman in 2014. EpiPALEOMIX encompasses diverse modules, among which the MethylMap module stands out, permitting users to derive methylation levels in regions that can be defined by the user (Hanghøj et al., 2016). However, this limitation is inherent to its usage; the user is required to have an understanding of the particular regions associated with the condition under study. The outcome of this algorithm is the calculated count of deaminated methylated cytosines in the CpG context and the corresponding coverage, representing their ratio, thereby denoting the methylation level at the particular genomic position. In contrast, the DamMet algorithm exhibits greater versatility. Unlike epiPALEOMIX, it is designed for whole-genome investigations. Furthermore, DamMet can calculate deamination levels in both methylated and unmethylated CpGs at each read position, thus employing a model that most accurately characterizes the deamination of cytosines in aDNA fragments as a random process (Hanghøj et al., 2019).

Regarding the handling of modern tissue samples, whole-genome bisulfite sequencing (WGBS) is the prevalent method for investigating DNA methylation (Olova et al., 2018; Suzuki et al., 2018). Several methods are available for reconstructing methylation from samples sequenced using this technology (Clark et al., 1994; Bock et al., 2005), with the most well-known being Bismark, BoostMe, and WGBStools. Currently, the Bismark algorithm is the most frequently used for preprocessing WGBS data. This involves the mapping of reads to the converted reference genome, followed by the quantification of methylated and unmethylated cytosines at each genomic position (Krueger, Andrews, 2011). Similar to many read-count-based methods, this approach is not well-suited for overcoming the challenge of low sample coverage, a common occurrence in cases involving low-quality samples or single-cell experiments. To address this concern, machine learning-based algorithms like DeepCPG and BoostMe have been created.

DeepCPG is a deep learning neural network-based algorithm designed to predict the methylation states of low-coverage sites and uncover motifs associated with changes in methylation levels and intercellular variability (Angermueller et al., 2017). This tool is primarily utilized to enhance the quality of data from single-cell experiments. BoostMe, which is based on a machine learning approach, addresses this issue during the genome preprocessing stage by employing imputation (Zou et al., 2018). The XGBoost gradient boosting technique employed in this tool amalgamates data from multiple samples (more than 3) to rectify missing methylation levels in contemporary tissue samples. This enables the utilization of low-coverage genome samples for methylation reconstruction. Additionally, a notable feature of BoostMe is its capacity to restore not only the state of a given CpG site (methylated/unmethylated) but also its methylation level. WGBStools, comprising a collection of methods developed in the context of the modern tissue methylation atlas project, is utilized for a highly efficient representation of mapped reads, statistical analysis, and visualization of data ranging from small genomic segments to entire chromosomal loci (https://github.com/nloyfer/wgbs_tools).

However, despite the variety of methylation reconstruction algorithms available, the application of WGBS technology to aDNA samples is limited. This limitation arises from the requirement for a high concentration of well-purified DNA for bisulfite conversion. Additionally, the bisulfite conversion process leads to DNA fragmentation, further compromising the quality of aDNA, which is already significantly fragmented due to degradation (Gu et al., 2011). Therefore, methylation level calculation algorithms commonly used for modern samples cannot be employed for the reconstruction of methylation profiles in ancient individuals. Consequently, our focus has been on developing a novel algorithm that enables the comparison of methylomes in ancient and modern samples, considering the lack of available bone tissue samples for conducting whole-genome bisulfite sequencing in the context of obesity.

Materials and methods

Sample selection. For our analysis, we curated a dataset from the NCBI GEO database, consisting of 11 ancient genomes and 12 modern methylation profiles obtained using Whole Genome Bisulfite Sequencing (WGBS) methods. When selecting the ancient samples, particular attention was given to the age of the samples, library preparation strategy, and genome coverage. We exclusively included samples that underwent prior USER treatment, were dated to be at least 3,000 years Before the Common Era (BCE), and had a minimum coverage of 5x. The complete genomes of ancient samples were sequenced with USER treatment, except for samples Vi33 and PES001 (Peschanitsa), which were not subjected to UDG treatment before sequencing (Table 1).

The selection of the 12 contemporary samples (Loyfer et al., 2023) was based on the mesodermal origin of the tissues used for library preparation, in conjunction with the utilization of whole-genome bisulfite sequencing. Additional information about these samples is presented in Table 2.

Ancient genomes preprocessing. The ancient genomes were obtained from ftp server in bam format along with their corresponding indices. As per previous studies (Ohm et al., 2010; Gokhman et al., 2014), it is well-recognized that UDG treatment is not sufficiently effective at the DNA termini. To ensure precise aDNA analysis, we employed the trimBam utility to trim two nucleotides from both the 3′ and 5′ ends of sequences (Gansauge, Meyer, 2013; Jun et al., 2015). It's important to note that for the Vi33 and PES001 samples, this trimming procedure was omitted due to the absence of UDG treatment during library preparation. Moreover, we applied Trimmomatic (Bolger et al., 2014) for the filtration of sample reads based on criteria such as average quality and length. In our subsequent analysis, only sequences that aligned with the CRCh37 (hg19) assembly and exhibited an average quality score exceeding 20, as well as a minimum length of 25 base pairs, were retained for further investigation.

Reconstruction of DNA methylation profiles in ancient humans. To reconstruct the methylation profiles of ancient samples, we utilized the DamMet software (Hanghøj et al., 2019). The pipeline consisted of three main stages: the filtration of single-nucleotide variants, the calculation of deamination levels for each read position, and the estimation of methylation levels.

The single-nucleotide variant (SNV) calling was performed using the GATK HaplotypeCaller v4.3.0.0 (Poplin et al.,

Table 1. Ancient genomes selected for analysis

Sample	Group	Sample age, kya	Sex	Tissue	Coverage	Methylation profile	Genomic smoothing window (CpG)	Reference
Altai Neanderthal	Ancient	120	Female	Toe phalanx	50	Gokhman et al., 2014, 2020	25	Prüfer et al., 2014
Denisovan		75	Female	Toe phalanx	30		25	Meyer et al., 2012
Vindija33		50	Female	Unknown bone	30		50	Prüfer et al., 2017
Ust'-Ishim	HG	45	Male	Femur	42 (22 XY)	Gokhman et al., 2020	25	Fu et al., 2014
Sunghir		35	Male	Femur + teeth	10.7	This study	38	Sikora et al., 2017
USR1		11.5	Female	Petrous bone	17		50	Moreno-Mayar et al., 2018a
Spirit Cave		11	Male	Petrous bone + teeth	18		33	Moreno-Mayar et al., 2018b
Peschanitsa		11	Male	Teeth	5		50	Saag et al., 2021
SF12		9	Female	Femur	57.79		28	Günther et al., 2018
2H10 (France)		3.2	Male	Teeth	13.9		33	Seguin-Orlando et al., 2021
2H11 (France)		3.2	Male	Teeth	23.9		33	Seguin-Orlando et al., 2021

Note. Smoothing window – a parameter for averaging deamination levels in the subsequent analysis stage. HG – hunter-gatherers.

Table 2. Contemporary genomes used for identifying tissue-specific methylated regions and DMRs

GEO accession	Sex	Age of patient	Organ	Tissue
GSM5652198	Male	37	Colon	Fibroblasts
GSM5652202	Female	35	Heart	
GSM5652204	Male	73	Derma	
GSM5652205	Female	59	Skeletal muscle	Smooth myocytes
GSM5652207	Male	22	Aorta	
GSM5652209	Female	51	Bladder	
GSM5652210	Male	24	Prostate	
GSM5652211	Male	57	Lung bronchus	
GSM5652212	Male	83	Heart	Cardiomyocytes
GSM2637888	–	–	Bone	–
GSM2637887	–	–	Bone	–
GSM5652218	Female	7	Bone	Osteoblasts
GSM5652177	Female	35	Subcutaneous adipose tissue	Adipocytes
GSM5652176	Female	53	Subcutaneous adipose tissue	
GSM5652178	Female	37	Subcutaneous adipose tissue	

2017). SNVs with coverage of less than 5 and quality less than 30 were filtered out. Additionally, variants were filtered when they exhibited homozygosity for the alternative allele or more than two alternative alleles when the position contained a cytosine. This stage followed the recommendations of the DamMet algorithm author, as described in Hanghøj et al., 2019, and supplementary materials provided therein.

Subsequently, methylation levels were reconstructed, excluding the identified variants.

```
DamMet estDEAM -b <bam-file> -r <fasta-file> -c
<chromosome> -M <expected-average-methylation>
-O <out-file-prefix> -E <vcf-to-exclude> -L 25
-P 50 -q 20 -Q 20
```

Subsequently, we determined the methylation levels based on the identified deamination levels at positions with both methylated and demethylated cytosines. The genomic window size for each sample is indicated in the respective column of Table 1 and was selected through empirical evaluation.

```
DamMet estF -b <bam-file> -r <fasta-file> -c
<chromosome> -M <expected-average-methylation>
-O <out-file-prefix> -N <genomic-window-size-
in-CpGs>
```

The acquired methylation profiles were additionally subjected to smoothing using a Python script that applied a moving average with a smoothing window size of 25 CpG sites.

Validation of the reconstructed methylomes. The comparison of Neanderthal, Denisovan, and Ust-Ishim hunter-gatherer methylomes obtained in the previous stage was conducted using the R programming language. We employed packages like ggplot, psych, corr.test, and the tidyverse family for data preprocessing, correlation analysis, and graph generation.

Identification of tissue-specific methylated regions. We designed a Python script for the identification of regions exhibiting relatively consistent methylation levels across all mesodermal tissues. This script takes the methylation values obtained using the Bismark algorithm (Krueger, Andrews, 2011) after aligning the aforementioned samples as input. It conducts a per-position comparison of methylation values through ANOVA to detect variations within three tissue groups (fibroblasts, myocytes, osteoblasts) and exclude positions showing statistically significant differential methylation ($p < 0.05$) from both ancient bone and modern adipocyte methylation profiles.

DMR identification. The prepared methylation profiles of hunter-gatherers (HG) and modern individuals were compared using the ANOVA method, similar to the tissue-specific methylation search. In the first iteration, the samples were divided into three groups: hunter-gatherer bone samples, healthy individuals' adipocytes, and obese patients' adipocytes. CpG sites with a significance level of $p < 0.05$ were selected for subsequent analysis using the Tukey *post hoc* test. A CpG site was considered differentially methylated if the methylation change was significant ($p < 0.05$) when comparing HG bones to adipocytes of obese individuals and not significant when comparing HG bones to controls.

In the second iteration, we modified the grouping: all samples were bone samples, and the groups represented samples of different ages (anatomically ancient humans, hunter-gatherers, and modern humans). Comparisons were made only in regions associated with obesity to reduce the computational load. To aggregate the obtained differentially methylated sites into regions, we used the combined-pvalues software

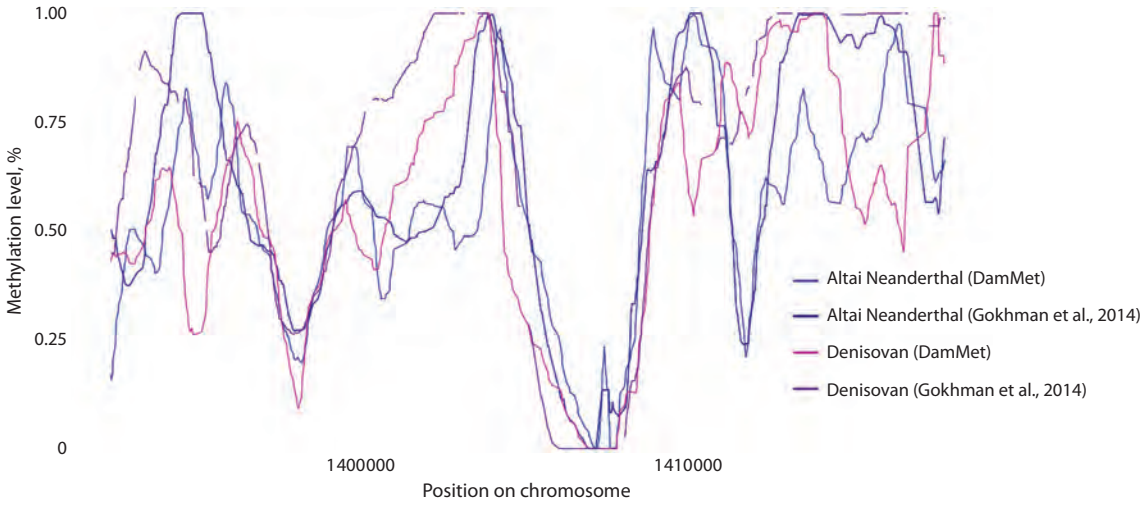


Fig. 1. Comparison of the methylation profiles of a Denisovan and a Neanderthal reconstructed by DamMet and published by D. Gokhman.

In focus: a demethylated CpG island at chr1:1406845–1407821.

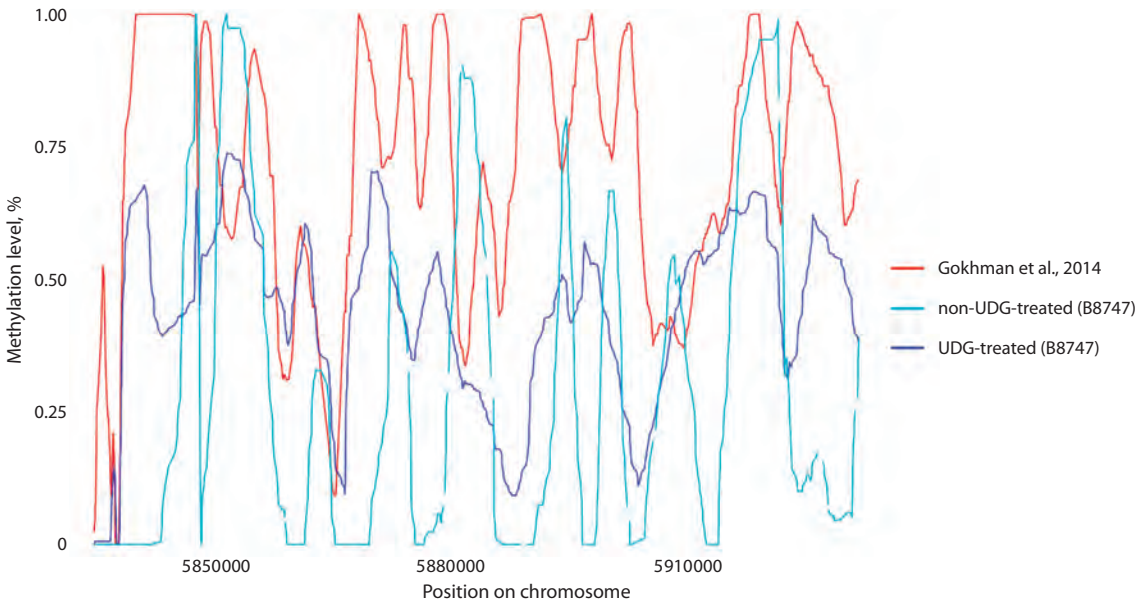


Fig. 2. Comparison of methylation levels on a region of chromosome 2 in sample Vi33, in the presence and absence of USER treatment during library preparation, with previously published profiles by D. Gokhman.

Methylation levels of all samples were smoothed using a 25 CpG moving average.

(https://github.com/brentp/combined-pvalues), which is based on the Stouffer–Liptak multiple testing correction method (Pedersen et al., 2012). The methylation change status was determined by comparing the mean methylation values in the regions between groups.

Results

In this study, we reconstructed 11 DNA methylation profiles of ancient humans using the DamMet tool. Firstly, we needed to develop a pipeline that would allow us to reconstruct methylomes with high precision. For this purpose, we used the genomes of Neanderthals and Denisovans, which had undergone UDG treatment, as input data for the pipeline. Profiles for these organisms had previously been published (Gokhman et al., 2014, 2020), enabling us to validate the pipeline. We found that our calculated methylation levels were, on average, 15–20 % lower than those previously published, but overall, the methylation profiles were similar (Fig. 1). The correlation coefficients for methylation profiles in both cases were over 85 %: $r_{\text{Denisovan}} = 0.87$, $r_{\text{Neanderthal}} = 0.9$ ($p < 0.05$).

As we had several samples that didn't undergo USER treatment during library preparation, we also aimed to confirm whether DamMet could reconstruct methylation profiles without this step. To address this, we selected sample Vi33, for which sequences both with and without USER treatment were

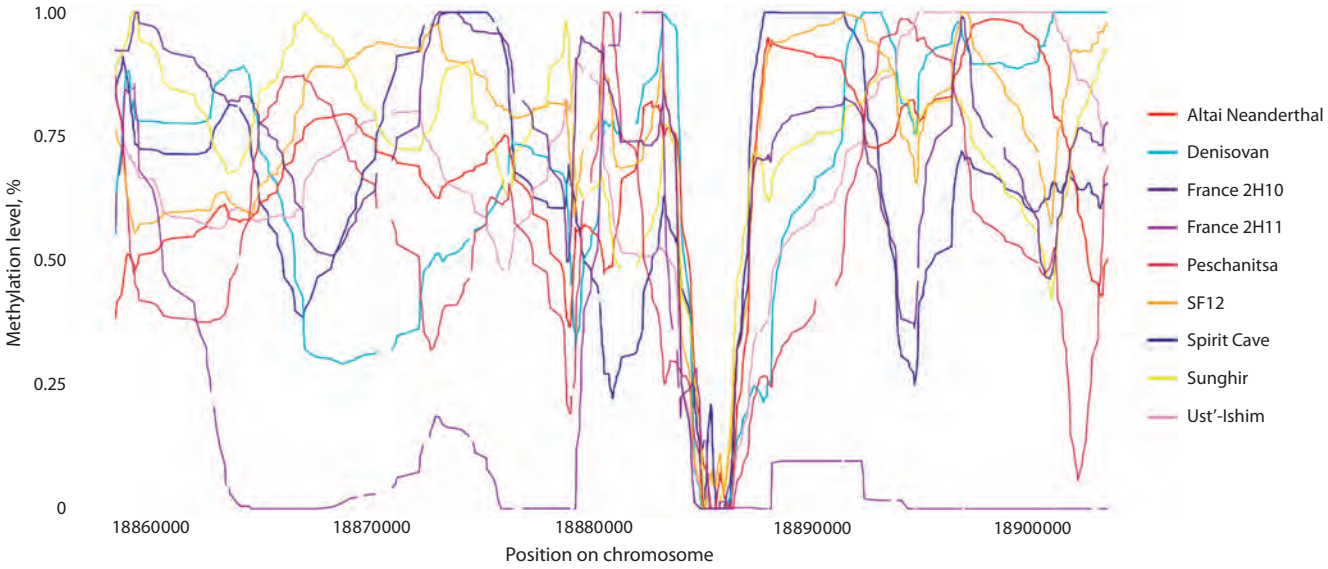


Fig. 3. Methylation profiles of hunter-gatherers reconstructed using DamMet.
The region of extensive demethylation corresponds to the CpG island at chr21:18884807–18886111 (GRCh37 hg19).

publicly available. The pipeline parameters were consistent for these analyses, ensuring uniform conditions for reconstructing methylomes from both libraries.

Our findings revealed that the methylation profile obtained in the presence of USER treatment showed an average correlation of 0.57 with the profile calculated by D. Gokhman, as depicted in Figure 2. In contrast, the methylome obtained without any treatment displayed a weak correlation ($r = 0.14$) with the published profile. Notably, the methylation patterns primarily matched in demethylated CpG islands, irrespective of whether we applied subsequent smoothing using a moving average.

Next, we processed eight genomes of hunter-gatherers using our pipeline, for which methylation profiles had not been reconstructed previously (see Table 1). The resulting profiles generally exhibited a similar methylation pattern to other ancient methylomes, including complete demethylation of some CpG islands (Fig. 3), resembling the profile of the previously reconstructed Ust-Ishim hunter-gatherer (Gokhman et al., 2020). Even though sample PES001 was not subjected to USER treatment during library preparation, our obtained methylation profile exhibited overall trends similar to other hunter-gatherer profiles and thus was not excluded from further analysis.

According to the authors of the method, the reconstructed methylation profiles using DamMet can be used for direct comparison with modern data. However, methylation can vary between cells of different origins, so direct comparisons should be limited to methylation profiles obtained from the same tissues. To the best of our knowledge, there has been no sequencing of bone tissues in the context of obesity. Therefore, for the final comparison, we selected samples from subcutaneous and visceral adipocyte tissues, which exhibit similar methylation patterns. However, these patterns may significantly differ from those observed in bones and other mesodermal tissues. As a result, we developed a Python script that performs a

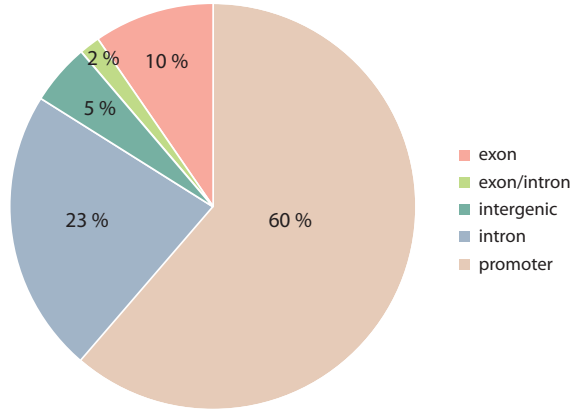


Fig. 4. Percentage distribution of DMRs in various genomic regions.

search for differentially methylated positions in mesodermal tissues and excludes them from further analysis. The script is based on dispersion analysis in three groups, followed by pairwise comparisons and multiple testing corrections. The mesodermal tissue samples were divided into groups according to tissue type: fibroblasts, muscle cells, and osteoblasts. In total, about 26.5 million CpG positions were analyzed, with approximately 206,000 showing differential methylation in at least one group, while more than 26 million did not exhibit significant differences.

We conducted a search for Differentially Methylated Regions (DMRs) in modern bone tissue samples, but focused our search on only 642 regions that had been previously associated with differential methylation in the context of obesity, as reported in the literature. In this case, we performed a per-position ANOVA analysis for groups of ancient individuals, hunter-gatherers, and modern individuals (bone tissue), with prior filtering of non-tissue-specific CpG sites. We identi-

fied 38 DMRs, where the overlap with the aforementioned 642 regions included more than 20 CpG sites. As depicted in Figure 4, approximately 60 % of these DMRs are located in gene promoter regions, 35 % are within gene body regions, and only 5 % of the DMRs are situated in intergenic regions. Notably, 94 % of these DMRs exhibit hypermethylation, potentially leading to the suppression of gene expression, particularly in genes associated with obesity.

Supplementary data and source code

The methylation profiles of ancient humans and the Python scripts used for the analysis in this study are available in the GitHub repository: https://github.com/bor-d/ancDMR

Conclusions

There are currently several methods available for reconstructing methylation profiles of ancient organisms, with epiPALEOMIX (Hanghøj et al., 2016) and DamMet (Hanghøj et al., 2019) being the two most commonly used ones. While both of these methods are known for their significant accuracy, their performance is often constrained by the quality of ancient DNA samples. In our study, we opted to utilize the DamMet method due to its versatility, specifically its capacity to compare the reconstructed methylation values with profiles generated using alternative sequencing technologies. However, during the validation of our pipeline, we observed notable discrepancies between the methylation values obtained with DamMet and those previously published by D. Gokhman, in both 2014 and 2020. The developers of DamMet acknowledge that their tool tends to yield lower methylation values in comparison to profiles generated using epiPALEOMIX, which does not account for factors such as single nucleotide variants (SNVs), sequencing errors, and the demethylation of unmethylated cytosines. This was evident in our reconstruction of Neanderthal and Denisovan profiles. Nonetheless, our analysis indicated a positive correlation between the methylation values reconstructed by DamMet and the previously published data. This reaffirms the tool's effectiveness in reconstructing previously uncharacterized methylation profiles, which can then be used for subsequent comparisons with modern methylomes.

In a demonstration of the pipeline we had devised, we attempted to identify DMRs within the genomic profiles of hunter-gatherers and contemporary humans, specifically in the context of obesity. We identified 38 regions, with approximately two-thirds of them located in promoter regions. This observation implies a plausible association between alterations in methylation patterns within these promoters and the regulation of gene expression. Certainly, the well-defined procedural stages within our pipeline effectively tackle potential hurdles researchers might face. This is especially valuable when dealing with situations where there is a lack of published methylation profiles related to the specific tissues of interest. These steps help reduce the likelihood of false-positive DMRs due to tissue-specificity.

When utilizing this pipeline to investigate DMRs related to different medical conditions, researchers are advised to conduct a thorough review of relevant scientific literature. This exploratory endeavour should ultimately lead to the discovery of regions where methylation patterns are inherently connected to the specific condition being studied. However, it is imperative to underscore that despite the explicit precautions taken, including the exclusion of tissue-specific regions and stringent filtering in the context of disease-associated regions, the investigation of DMRs may still encompass CpG sites, the methylation profiles of which underwent alterations during the evolutionary transition from archaic humans (*Homo sapiens neanderthalensis*) to contemporary *Homo sapiens sapiens*.

References

Angermueller C., Lee H.J., Reik W., Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol*. 2017;18(1):67. DOI 10.1186/s13059-017-1189-z

Bock C., Reither S., Mikeska T., Paulsen M., Walter J., Lengauer T. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*. 2005;21(21): 4067-4068. DOI 10.1093/bioinformatics/bti652

Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. DOI 10.1093/bioinformatics/btu170

Briggs A.W., Stenzel U., Johnson P.L.F., Green R.E., Kelso J., Prüfer K., Meyer M., Krause J., Ronan M.T., Lachmann M., Pääbo S. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. USA*. 2007;104(37):14616-14621. DOI 10.1073/pnas.0704665104

Briggs A.W., Good J.M., Green R.E., Krause J., Maricic T., Stenzel U., Lalueza-Fox C., Rudan P., Brajković D., Kućan Ž., Gušić I., Schmitz R., Doronichev V.B., Golovanova L.V., de la Rasilla M., Fortea J., Rosas A., Pääbo S. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*. 2009a;325(5938):318-321. DOI 10.1126/science.1174462

Briggs A.W., Good J.M., Green R.E., Krause J., Maricic T., Stenzel U., Pääbo S. Primer extension capture: targeted sequence retrieval from heavily degraded DNA sources. *J. Vis. Exp*. 2009b;31:1573. DOI 10.3791/1573

Briggs A.W., Stenzel U., Meyer M., Krause J., Kircher M., Pääbo S. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res*. 2010;38(6):e87. DOI 10.1093/nar/gkp1163

Clark S.J., Harrison J., Paul C.L., Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*. 1994;22(15):2990-2997. DOI 10.1093/nar/22.15.2990

Feinberg A.P., Irizarry R.A. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. USA*. 2010;107(Suppl.1):1757-1764. DOI 10.1073/pnas.0906183107

Fu Q., Li H., Moorjani P., Jay F., Slepchenko S.M., Bondarev A.A., Johnson P.L.F., Aximu-Petri A., Prüfer K., de Filippo C., Meyer M., Zwyns N., Salazar-García D.C., Kuzmin Y.V., Keates S.G., Kosintsev P.A., Razhev D.I., Richards M.P., Peristov N.V., Lachmann M., Douka K., Higham T.F.G., Slatkin M., Hublin J.J., Reich D., Kelso J., Viola T.B., Pääbo S. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. 2014;514(7523):445-449. DOI 10.1038/nature13810

Gansauge M.-T., Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc*. 2013; 8(4):737-748. DOI 10.1038/nprot.2013.038

Gokhman D., Lavi E., Prüfer K., Fraga M.F., Riancho J.A., Kelso J., Pääbo S., Meshorer E., Carmel L. Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*. 2014; 344(6183):523-527. DOI 10.1126/science.1250368

Gokhman D., Nissim-Rafinia M., Agranat-Tamir L., Housman G., García-Pérez R., Lizano E., Cheronet O., Mallick S., Nieves-Colón M.A., Li H., Alpaslan-Roodenberg S., Novak M., Gu H., Osinski J.M., Fer-

rando-Bernal M., Gelabert P., Lipende I., Mjungu D., Kondova I., Bontrop R., Kullmer O., Weber G., Shahar T., Dvir-Ginzberg M., Faerman M., Quillen E.E., Meissner A., Lahav Y., Kandel L., Liebergall M., Prada M.E., Vidal J.M., Gronostajski R.M., Stone A.C., Yakir B., Lalueza-Fox C., Pinhasi R., Reich D., Marques-Bonet T., Meshorer E., Carmel L. Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun*. 2020; 11(1):1189. DOI 10.1038/s41467-020-15020-6

Gu H., Smith Z.D., Bock C., Boyle P., Gnirke A., Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc*. 2011;6(4): 468-481. DOI 10.1038/nprot.2010.190

Günther T., Malmström H., Svensson E.M., Omrak A., Sánchez-Quinto F., Kılınç G.M., Krzewińska M., Eriksson G., Fraser M., Edlund H., Munters A.R., Coutinho A., Simões L.G., Vicente M., Sjölander A., Sellevold B.J., Jørgensen R., Claes P., Shriver M.D., Valdiosera C., Netea M.G., Apel J., Lidén K., Skar B., Storå J., Götherström A., Jakobsson M. Population genomics of Mesolithic Scandinavia: investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol*. 2018;16(1):e2003703. DOI 10.1371/journal.pbio.2003703

Hanghøj K., Seguin-Orlando A., Schubert M., Madsen T., Pedersen J.S., Willerslev E., Orlando L. Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Mol. Biol. Evol*. 2016;33(12):3284-3298. DOI 10.1093/molbev/msw184

Hanghøj K., Renaud G., Albrechtsen A., Orlando L. DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *GigaScience*. 2019;8(4):giz025. DOI 10.1093/gigascience/giz025

Jablonka E., Raz G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol*. 2009;84(2):131-176. DOI 10.1086/598822

Jun G., Wing M.K., Abecasis G.R., Kang H.M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015;25(6): 918-925. DOI 10.1101/gr.176552.114

Krueger F., Andrews S. RBismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27(11): 1571-1572. DOI 10.1093/bioinformatics/btr167

Loyfer N., Magenheim J., Peretz A., Cann G., Bredno J., Klochendler A., Fox-Fisher I., Shabi-Porat S., Hecht M., Pelet T., Moss J., Drawshy Z., Amini H., Moradi P., Nagaraju S., Bauman D., Shveiky D., Porat S., Dior U., Rivkin G., Or O., Hirshoren N., Carmon E., Pikarsky A., Khalaileh A., Zamir G., Grinbaum R., Gazala M.A., Mizrahi I., Shussman N., Korach A., Wald O., Izhar U., Erez E., Yutkin V., Samet Y., Golinkin D.R., Spalding K.L., Druid H., Arner P., Shapiro A.M.J., Grompe M., Aravanis A., Venn O., Jamshidi A., Shemer R., Dor Y., Glaser B., Kaplan T. A DNA methylation atlas of normal human cell types. *Nature*. 2023;613(7943):355-364. DOI 10.1038/s41586-022-05580-6

Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., de Filippo C., Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. A high-coverage genome sequence from an archaic Denisovan individual. *Science*. 2012;338(6104):222-226. DOI 10.1126/science.1224344

Moreno-Mayar J., Potter B., Vinner L., Steinrücken M., Rasmussen S., Terhorst J., Kamm J., Albrechtsen A., Malaspinas A., Sikora M., Reuther J., Irish J., Malhi R., Orlando L., Song Y., Nielsen R., Meltzer D., Willerslev E. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*. 2018a; 553(7687):203-207. DOI 10.1038/nature25173

Moreno-Mayar J.V., Vinner L., Damgaard P.B., de la Fuente C., Chan J., Spence J.P., Allentoft M.E., Vimala T., Racimo F., Pinotti T., Rasmussen S., Margaryan A., Orbegozo M.I., Mylopotamitaki D., Wooller M., Bataille C., Becerra-Valdivia L., Chivall D., Comeskey D., Devièse T., Grayson D.K., George L., Harry H., Alexandersen V., Primeau C., Erlandson J., Rodrigues-Carvalho C., Reis S., Bastos M.Q.R., Cybulski J., Vullo C., Morello F., Vilar M., Wells S., Gregersen K., Hansen K.L., Lynnerup N., Mirazón Lahr M., Kjær K., Strauss A., Alfonso-Durruty M., Salas A., Schroeder H., Higham T., Malhi R.S., Rasic J.T., Souza L., Santos F.R., Malaspinas A.-S., Sikora M., Nielsen R., Song Y.S., Meltzer D.J., Willerslev E. Early human dispersals within the Americas. *Science*. 2018b;362(6419). DOI 10.1126/science.aav2621

Niiranen L., Leciej D., Edlund H., Bernhardsson C., Fraser M., Sánchez Quinto F., Herzig K.H., Jakobsson M., Walkowiak J., Thalmann O. Epigenomic modifications in modern and ancient genomes. *Genes*. 2022;13(2):178. DOI 10.3390/genes13020178

Ohm J.E., Mali P., Van Neste L., Berman D.M., Liang L., Pandiyan K., Briggs K.J., Zhang W., Argani P., Simons B., Yu W., Matsui W., Van Criekinge W., Rassool F.V., Zambidis E., Schuebel K.E., Cope L., Yen J., Mohammad H.P., Cheng L., Baylin S.B. Cancer-related epigenome changes associated with reprogramming to induced pluripotent stem cells. *Cancer Res*. 2010;70(19):7662-7673. DOI 10.1158/0008-5472.CAN-10-1361

Olova N., Krueger F., Andrews S., Oxley D., Berrens R.V., Branco M.R., Reik W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol*. 2018;19(1):33. DOI 10.1186/s13059-018-1408-2

Orlando L., Gilbert M.T.P., Willerslev E. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet*. 2015;16(7):395-408. DOI 10.1038/nrg3935

Pedersen B.S., Schwartz D.A., Yang I.V., Kechris K.J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *P*-values. *Bioinformatics*. 2012;28(22):2986-2988. DOI 10.1093/bioinformatics/bts545

Poplin R., Ruano-Rubio V., DePristo M.A., Fennell T.J., Carneiro M.O., Van der Auwera G.A., Kling D.E., Gauthier L.D., Levy-Moonshine A., Roazen D., Shakir K., Thibault J., Chandran S., Whelan C., Lek M., Gabriel S., Daly M.J., Neale B., MacArthur D.G., Banks E. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2017. DOI 10.1101/201178

Prüfer K., Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P.H., de Filippo C., Li H., Mallick S., Dannemann M., Fu Q., Kircher M., Kuhlwilm M., Lachmann M., Meyer M., Ongyerth M., Siebauer M., Theunert C., Tandon A., Moorjani P., Pickrell J., Mullikin J.C., Vohr S.H., Green R.E., Hellmann I., Blanche H., Cann H., Kitzman J.O., Shendure J., Eichler E.E., Lein E.S., Bakken T.E., Golovanova L.V., Doronichev V.B., Shunkov M.V., Derevianko A.P., Viola B., Slatkin M., Reich D., Kelso J., Pääbo S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 2014;505(7481): 43-49. DOI 10.1038/nature12886

Prüfer K., de Filippo C., Grote S., Mafessoni F., Korlević P., Hajdinjak M., Vernot B., Skov L., Hsieh P., Peyrégne S., Reher D., Hopfe C., Nagel S., Maricic T., Fu Q., Theunert C., Rogers R., Skoglund P., Chintalapati M., Dannemann B., Nelson B.J., Key F.M., Rudan P., Kućan Ž., Gušić I., Golovanova L.V., Doronichev V.B., Patterson N., Reich D., Eichler E.E., Slatkin M., Schierup M.H., Andrés A.M., Kelso J., Meyer M., Pääbo S. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*. 2017;358(6363):655-658. DOI 10.1126/science.aao1887

Saag L., Vasilyev S.V., Varul L., Kosorukova N.V., Gerasimov D.V., Oshibkina S.V., Griffith S.J., Solnik A., Saag L., D'Atanasio E., Metspalu E., Reidla M., Rootsi S., Kivisild T., Scheib C.L., Tambets K., Kriiska A., Metspalu M. Genetic ancestry changes in Stone to Bronze Age transition in the East European plain. *Sci. Adv*. 2021;7:eabd6535. DOI 10.1126/sciadv.abd6535

Sawyer S., Krause J., Guschanski K., Savolainen V., Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in

ancient DNA. *PLoS One*. 2012;7(3):e34131. DOI 10.1371/journal.pone.0034131

Seguin-Orlando A., Donat R., Der Sarkissian C., Southon J., Thèves C., Manen C., Tchérémissinoff Y., Crubézy E., Shapiro B., Deleuze J., Dalén L., Guilaine J., Orlando L. Heterogeneous hunter-gatherer and steppe-related ancestries in Late Neolithic and Bell Beaker genomes from present-day France. *Curr. Biol.* 2021;31(5):1072-1083. DOI 10.1016/j.cub.2020.12.015

Sikora M., Seguin-Orlando A., Sousa V.C., Albrechtsen A., Korneliussen T., Ko A., Rasmussen S., Dupanloup I., Nigst P.R., Bosch M.D., Renaud G., Allentoft M.E., Margaryan A., Vasilyev S.V., Veselovskaya E.V., Borutskaya S.B., Deviese T., Comeskey D., Higham T., Manica A., Foley R., Meltzer D.J., Nielsen R., Excoffier L., Lahr M.M., Orlando L., Willerslev E. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*. 2017;358(6363):659-662. DOI 10.1126/science.aao1807

Suzuki M., Liao W., Wos F., Johnston A.D., DeGrazia J., Ishii J., Bloom T., Zody M.C., Germer S., Greally J.M. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res.* 2018;28(9):1364-1371. DOI 10.1101/gr.232587.117

Zhur K.V., Trifonov V.A., Prokhortchouk E.B. Progress and prospects in epigenetic studies of ancient DNA. *Biochemistry* (*Mosc.*). 2021; 86(12-13):1563-1571. DOI 10.1134/S0006297921120051

Zou L.S., Erdos M.R., Taylor D.L., Chines P.S., Varshney A., Parker S.C.J., Collins F.S., Didion J.P. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics*. 2018;19(1):390. DOI 10.1186/s12864-018-4766-y

ORCID ID

D.D. Borodko orcid.org/0000-0003-3596-5470
S.V. Zhenilo orcid.org/0000-0003-0874-1594
F.S. Sharko orcid.org/0000-0002-1189-5597

Acknowledgements. The work was carried out with financial support from the project of the Ministry of Science and Higher Education of the Russian Federation, under grant number 075-10-2020-116 (grant ID 13.1902.21.0023).

Conflict of interest. The authors declare no conflict of interest.

Received July 15, 2023. Revised October 4, 2023. Accepted October 5, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):829-838
DOI 10.18699/VJGB-23-96

Original Russian text https://vavilovj-icg.ru/

Evolution of human genes encoding cell surface receptors involved in the regulation of appetite: an analysis based on the phylostratigraphic age and divergence indexes

E.V. Ignatieva ✉, S.A. Lashin, Z.S. Mustafin, N.A. Kolchanov

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ eignat@bionet.nsc.ru

Abstract. Genes encoding cell surface receptors make up a significant portion of the human genome (more than a thousand genes) and play an important role in gene networks. Cell surface receptors are transmembrane proteins that interact with molecules (ligands) located outside the cell. This interaction activates signal transduction pathways in the cell. A large number of exogenous ligands of various origins, including drugs, are known for cell surface receptors, which accounts for interest in them from biomedical researchers. Appetite (the desire of the animal organism to consume food) is one of the most primitive instincts that contribute to survival. However, when the supply of nutrients is stable, the mechanism of adaptation to adverse factors acquired in the course of evolution turned out to be excessive, and therefore obesity has become one of the most serious public health problems of the twenty-first century. Pathological human conditions characterized by appetite violations include both hyperphagia, which inevitably leads to obesity, and anorexia nervosa induced by psychosocial stimuli, as well as decreased appetite caused by neurodegeneration, inflammation or cancer. Understanding the evolutionary mechanisms of human diseases, especially those related to lifestyle changes that have occurred over the past 100–200 years, is of fundamental and applied importance. It is also very important to identify relationships between the evolutionary characteristics of genes in gene networks and the resistance of these networks to changes caused by mutations. The aim of the current study is to identify the distinctive features of human genes encoding cell surface receptors involved in appetite regulation using the phylostratigraphic age index (PAI) and divergence index (DI). The values of PAI and DI were analyzed for 64 human genes encoding cell surface receptors, the orthologs of which were involved in the regulation of appetite in model animal species. It turned out that the set of genes under consideration contains an increased number of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence), and almost all of these genes (28 out of 31) belong to the superfamily of G-protein coupled receptors. Apparently, the synchronized evolution of such a large group of genes (31 genes out of 64) is associated with the development of the brain as a separate organ in the first vertebrates. When studying the distribution of genes from the same set by DI values, a significant enrichment with genes having a low DIs was revealed: eight genes (*GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1*) had extremely low DIs (less than 0.05). Such low DI values indicate that most likely these genes are subjected to stabilizing selection. It was also found that the group of genes with low DIs was enriched with genes that had brain-specific patterns of expression. In particular, *GPR26*, which had the lowest DI, is in the group of brain-specific genes. Because the endogenous ligand for the GPR26 receptor has not yet been identified, this gene seems to be an extremely interesting object for further theoretical and experimental research. We believe that the features of the genes encoding cell surface receptors we have identified using the evolutionary metrics PAI and DI can be a starting point for further evolutionary analysis of the gene network regulating appetite.

Key words: regulation of appetite; cell surface receptors; hunger; evolution; phylostratigraphic analysis; gene age; gene variability.

For citation: Ignatieva E.V., Lashin S.A., Mustafin Z.S., Kolchanov N.A. Evolution of human genes encoding cell surface receptors involved in the regulation of appetite: an analysis based on the phylostratigraphic age and divergence indexes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):829-838. DOI 10.18699/VJGB-23-96

Анализ особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, на основе индексов филостратиграфического возраста и микроэволюционной изменчивости

Е.В. Игнатьева ✉, С.А. Лашин, З.С. Мустафин, Н.А. Колчанов

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ eignat@bionet.nsc.ru

Аннотация. Гены рецепторов клеточной поверхности составляют существенную долю генома человека (более тысячи генов) и выполняют важную роль в генных сетях. Рецепторы клеточной поверхности – это трансмембранные белки, которые взаимодействуют с различными молекулами (лигандами), находящимися во внеклеточном пространстве, что приводит к активации путей сигнальной трансдукции в клетке. Для рецепторов клеточной поверхности известно большое количество экзогенных лигандов различного происхождения, включая лекарственные препараты, что и определяет интерес к их исследованию с точки зрения биомедицины. Аппетит (стремление животного организма потреблять пищу) – один из самых примитивных инстинктов, способствующих выживанию. Однако приобретенный в ходе эволюции механизм приспособления к неблагоприятным факторам в условиях стабильного поступления питательных веществ оказался избыточным, в связи с чем ожирение стало одной из самых серьезных проблем общественного здравоохранения в XXI веке. Патологические состояния человека, характеризующиеся нарушениями аппетита, включают как гиперфагию, неминуемо приводящую к ожирению, так и нервную анорексию, индуцированную психосоциальными стимулами, и снижение аппетита, связанное с воспалительными, нейродегенеративными и онкологическими заболеваниями. Понимание эволюционных механизмов развития болезней человека, особенно связанных с изменениями образа жизни, произошедшими в течение последних 100–200 лет, имеет как фундаментальное, так и прикладное значение. Особенно важно установить взаимосвязи между эволюционными характеристиками генов в генных сетях и устойчивостью этих сетей к изменениям, вызванным мутациями. Цель данной работы – выявление особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, с использованием филостратиграфического индекса PAI (phylostratigraphic age index) и индекса эволюционной изменчивости DI (divergence index). Были проанализированы индексы PAI и DI для 64 генов человека, кодирующих рецепторы клеточной поверхности, ортологи которых участвовали в регуляции аппетита у модельных видов животных. Оказалось, что в рассматриваемом наборе генов содержится повышенное количество генов, имеющих одинаковый филостратиграфический возраст (PAI = 5, этап дивергенции позвоночных), и почти все эти гены (28 из 31) относятся к суперсемейству рецепторов, сопряженных с G-белком. По-видимому, синхронизированное эволюционирование такой многочисленной группы генов (31 из 64 генов) связано с формированием у первых позвоночных мозга как отдельного органа. При исследовании распределения генов из этого же набора по значениям индексов DI была выявлена существенная обогащенность генами с низким DI. При этом восемь генов (*GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1*) характеризовались экстремально низким значением DI (менее 0.05), что указывает на существенную их подверженность стабилизирующему отбору. Обнаружено также, что группа генов с низким DI обогащена генами, тканеспецифически экспрессирующимися в мозге. В частности, к группе генов, тканеспецифически экспрессирующихся в мозге, относится *GPR26*, имеющий самое низкое значение DI. Ввиду того, что эндогенный лиганд для рецептора GPR26 пока не выявлен, этот ген представляется чрезвычайно интересным объектом для дальнейшего теоретического и экспериментального исследования. Выявленные нами особенности распределения генов рецепторов клеточной поверхности по эволюционным индексам PAI и DI являются отправной точкой для дальнейшего анализа эволюционных характеристик генной сети регуляции аппетита в целом.

Ключевые слова: регуляция аппетита; рецепторы клеточной поверхности; чувство голода; эволюция; филостратиграфия; возраст гена; изменчивость генов.

Introduction

Appetite (the desire of the animal organism to consume food) is a physiological mechanism (feeling) that regulates the intake of nutrients. The desire to consume food is one of the most primitive instincts that contribute to survival. This instinct has been formed over millions of years of evolution of living beings and has provided powerful mechanisms for adaptation and response to periods of nutrient shortage (Yeo, Heisler, 2012). The ability to consume excessive amounts of food during periods of its availability significantly affected the survival of individuals both in human populations and in populations of other animal species.

With the development of human civilization, the human populations living in developed countries faced the problem of adaptation to the abundance of food combined with a decrease in physical activity, making obesity one of the most serious public health problems of the twenty-first century (Kaidar-Person et al., 2011). Thus, the mechanism of adaptation to

unfavourable factors acquired during evolution in conditions of stable nutrient supply turned out to be excessive (Yeo, Heisler, 2012).

In humans and other animal species, the physiological system that regulates appetite functions with the participation of protein products of genes expressed both in the brain (Olszewski et al., 2008) and in peripheral organs and tissues: stomach, intestine, pancreas, adipose tissue. Neurons involved in the regulation of the motivational drive to obtain food are located in different parts of the brain (hypothalamic nuclei, amygdala, dorsal raphe nucleus, nuclei of the solitary tract, ventral tegmental area, prefrontal cortex, etc.). They integrate signals received from the sensory organs (olfactory, visual, taste sensations) as well as various interoceptive and humoral signals and control search for food and food consumption (Yeo, Heisler, 2012; Tremblay, Bellisle, 2015; Heisler, Lam, 2017).

Appetite can be induced by energy and nutrient shortages (in this case it is referred to by the term homeostatic appetite). However, even in the absence of apparent homeostatic needs, factors such as the sight, smell and taste of food, environmental cues, and the anticipation of new sensations that arise from eating can stimulate eating behavior, i. e. non-homeostatic appetite. The neuronal systems controlling homeostatic and non-homeostatic appetite function in close cooperation (Ahn et al., 2022).

Neurons of the arcuate nucleus of the hypothalamus secreting neuropeptide Y (NPY), agouti-like protein (AgRP), and alpha-melanocyte stimulating hormone (α-MSH), which is generated as a proteolytic cleavage product from proopiomelanocortin (POMC) by prohormone convertases (PCSK1 and PCSK2), are central to the systems regulating both homeostatic and non-homeostatic appetite (Yeo, Heisler, 2012). The activity of neurons located in the arcuate nucleus is controlled by hormones (leptin, insulin, ghrelin, polypeptide YY (PYY), glucocorticoids, adrenocorticotropin, corticotropin-releasing hormone), neurotransmitters (serotonin, dopamine, adrenaline, GABA), and neurotrophic factors (BDNF, etc.) as well (Maniam, Morris, 2012; Yeo, Heisler, 2012; Heisler, Lam, 2017).

Human pathological conditions characterized by appetite disorders are known. A pathological increase in body weight (obesity) can be caused by such a condition as hyperphagia (bulimia). A catastrophic decrease in appetite is seen in anorexia nervosa, which is extremely dangerous and increases the risk of death in young people tenfold (Fichter, Quadflieg, 2016). Reduced appetite can accompany chronic inflammatory and autoimmune processes, cancer and neurodegenerative diseases (Grossberg et al., 2010). In this context, any new knowledge about the system of genes regulating appetite is of particular importance.

Previously, we performed a functional analysis of genes involved in the regulation of appetite and body weight (Ignatieva et al., 2014, 2016). When analyzing a set of 105 genes involved in appetite regulation, a statistically significant over-enrichment of genes specifically expressed in the brain was found. It was also revealed that a substantial proportion of genes (~45 %) in this set were genes encoding cell surface receptors. Many of these receptors belonged to the superfamily of G-protein-coupled receptors (GPCRs).

The GPCRs superfamily includes proteins that have a similar structure (all of them contain 7 transmembrane domains). These proteins can be found on the cell membranes of almost all eukaryotes (New, Wong, 1998; Yang et al., 2021). Analysis of the DNA sequence of the human genome made it possible to predict about 800 genes encoding proteins of this superfamily (including 388 genes encoding olfactory receptors) (Bjarnadóttir et al., 2006). GPCRs mediate the response of cells to extracellular signaling molecules of different nature – proteins, peptides, low molecular weight substances (odorous and taste stimuli, hormones), as well as light-sensitive compounds. In turn, these receptors activate signal transduction pathways in cells, providing fundamental physiological processes (vision, perception of taste and olfactory signals, neuronal functioning, endocrine regulation and reproduction processes) (Katritch et al., 2013). Some of the best known receptors from the GPCR superfamily, which we have previously classified as appetite-regulating genes (Igatieva et al., 2016), include, for example, GHSR (growth hormone secretagogue receptor), MC3R (melanocortin 3 receptor), MC4R (melanocortin 4 receptor), CCKAR (cholecystokinin A receptor), CCKBR (cholecystokinin B receptor) and GCGR (glucagon).

Understanding the evolutionary mechanisms of human diseases, especially those associated with lifestyle changes that have occurred over the last 100–200 years (and the above-mentioned diseases associated with appetite dysregulation are just such diseases), is of great fundamental and applied significance. It is also very important to find interdependence between the evolutionary characteristics of genes in gene networks and the resistance of these networks to disruptions of genes themselves (through mutations) and to alterations in gene expression patterns caused by genetic variability of regulatory regions. Phylogenetic and population analysis of genes and gene networks involved in the relevant biological processes may be useful in developing new scenarios for personalized prevention and targeted drug therapy of diseases.

The aim of this work was to identify the evolutionary features of human cell surface receptor genes involved in appetite regulation using phylostratigraphic age index (PAI) and divergence index (DI). To achieve this goal, at the first stage, a set of human receptor genes the orthologues of which were involved in appetite regulation in model organisms was formed based on the analysis of scientific publications. Next, the distributions of human genes according to PAI and DI values were examined. The characteristic features of these distributions were identified by comparison with the distributions obtained for all human protein-coding genes, as well as for genes encoding GPCRs.

Materials and methods

Collecting the list of genes involved in appetite regulation and encoding cell surface receptors. The list of genes was taken from (Ignatieva et al., 2016) and expanded based on a PubMed search (https://pubmed.ncbi.nlm.nih.gov/) using the keywords listed in Supplementary Material 1[1]. Only genes from experimental studies were considered; reviews were excluded. In most studies, the role of genes in the

[1] Supplementary Materials 1–13 are available at: https://vavilov.elpub.ru/jour/manager/files/Suppl_Ignatieva_Engl_27_7.pdf

Table 1. Gene sets for which the distributions of PAI and DI values were analysed

Short name	Description	Number of genes
allCDS_19,566	All protein-coding genes of the human genome for which PAI and DI values were calculated	19,566
Receptors_64	Human genes encoding cell surface receptors and involved in appetite regulation*	64
allGPCR_389	Human genes encoding GPCRs (this set included genes from the the GPCRdb (https://gpcrdb.org) with the exception of genes encoding olfactory receptors)	389
appGPCR_45	Genes from the *Receptors_64* set that encode GPCRs	45
app_not_GPCR_19	Genes from the *Receptors_64* set encoding receptors that do not belong to the G-protein-coupled receptor superfamily	19

* This set includes human genes orthologous to genes of other animal species, the role of which in appetite regulation has been studied experimentally.

regulation of food consumption was established using model organisms (mice, rats, etc.). Therefore, the list of human genes controlling appetite included orthologues of those genes that were identified in experiments on other animal species. Indication that the product of a gene is a cell surface receptor was obtained from the text field "Summary" of the EntrezGene database (https://www.ncbi.nlm.nih.gov/gene).

Control sets of genes. The human gene sets listed in Table 1 were also used in the analyses. The list of human genes encoding receptors and controlling appetite was named *Receptors_64*.

The set containing all human protein-coding genes (*allCDS_19,566*) included 19,566 protein-coding genes for which PAI and DI values were identified.

The set containing human genes encoding GPCRs (*allGPCR_389*) was formed on the basis of the GPCRdb (https://gpcrdb.org) (Pandy-Szekeres et al., 2023). Genes encoding olfactory receptors were not included in this set because the set of appetite-controlling cell surface receptor genes (*Receptors_64*) did not contain genes encoding receptors of this type.

The set containing genes encoding GPCRs that control appetite (*appGPCR_45*) was obtained by the intersection of two sets – *Receptors_64* and *allGPCR_389*.

Analysis of the evolutionary characteristics of genes. The analysis was performed using PAI (phylostratigraphic age index) and DI (divergence index).

Phylostratigraphic age index (PAI) shows to what extent the taxon reflecting the age of the gene is distant from the root of the phylogenetic tree. The taxon reflecting the age of the gene is understood as the taxon at which the divergence of the studied species from the most distant related taxon, in which the orthologue of the gene in question was found, occurred (Table 2). The greater the PAI value of the gene under study, the younger this gene is supposed to be. PAI values were calculated in the Orthoscape tool based on the KEGG Orthology service as described in (Mustafin et al., 2021). We used PAI values calculated at a similarity level of 0.5.

Divergence index (DI) is an index of evolutionary variability of a gene. DI is calculated based on the dN/dS ratio, where dN is the proportion of nonsynonymous substitutions in DNA sequences of the studied gene and its orthologue; dS is the proportion of synonymous substitutions. The value of this index was calculated based on the comparison of human genes with genes of closely related organisms from the Hominidae family, as described in (Mustafin et al., 2021). Thus, DI can be determined only for protein-coding genes and indicates the type of selection acting on the gene. DI value in the range from 0 to 1 shows that the gene is subjected to stabilizing selection, 1 – to neutral evolution, and more than 1 – to driving selection.

Table 2. PAI values and taxonomic units dating the corresponding phylostratigraphic age of genes

PAI	Taxonomic unit
0	Cellular Organisms
1	Eukaryota
2	Metazoa
3	Chordata
4	Craniata
5	Vertebrata
6	Euteleostomi
7	Mammalia
8	Eutheria
9	Euarchontoglires
10	Primates
11	Haplorrhini
12	Catarrhini
13	Hominidae
14	Homo
15	Homo sapiens

Analysis of tissue-specific characteristics of genes. We used the TSEA tool to find overrepresented groups of genes that had tissue-specific expression patterns identified for a certain organ or tissue (Wells et al., 2015). The TSEA tool (http://genetics.wustl.edu/jdlab/tsea/) uses data on tissue-specific gene expression patterns detected in 25 different human organs and tissues. The TSEA tool identifies groups of tissue-specific genes in the analyzed list of genes, the size of which

significantly exceeds the expected one for random reasons. The TSEA tool uses data on specificity indices (SI) of gene expression products and their corresponding *p*-values (pSI). These values were calculated for each organ or tissue and for each transcript based on the analysis of data obtained by whole transcriptome profiling (GTEx Consortium, 2015). If the pSI value was < 0.01, the transcript was considered as tissue-specific for a given tissue.

Statistical analysis. The significance of differences between the observed and expected numbers of genes in subgroups was evaluated using the Chi-Square test.

Results

Genes encoding cell surface receptors and their functional characteristics

As a result of queries to PubMed, experimental data on genes of model organisms (mice, rats, etc.) involved in the regulation of food intake were found. Using this information, as well as information from EntrezGene, we found 64 human genes orthologous to genes identified in model organisms and encoding cell surface receptors (in Table 1 this set of genes is presented as *Receptors_64*). See Supplementary Material 2 for the full list of genes.

We compared these 64 genes with those accumulated in GPCRdb and found that 45 (70.3 %) out of 64 genes encoded G-protein-coupled receptors (Fig. 1, *a*). As is shown in Table 1, this subset of genes encoding receptors from the GPCR superfamily was named *appGPCR_45* (Supplementary Material 2 contains data on whether the gene belongs to the GPCR superfamily). The remaining 19 genes (29.7 %) encoded receptors from the other superfamilies (this subset is represented in Table 1 as *app_not_GPCR_19*).

Analysis of the gene list using the TSEA tool (Wells et al., 2015) revealed that the *Receptors_64* set was enriched in genes that have brain-specific expression pattern. Approximately one fifth of the genes (12 genes, or 18.75 %) fall into this category (see Fig. 1, *b*, Supplementary Material 3).

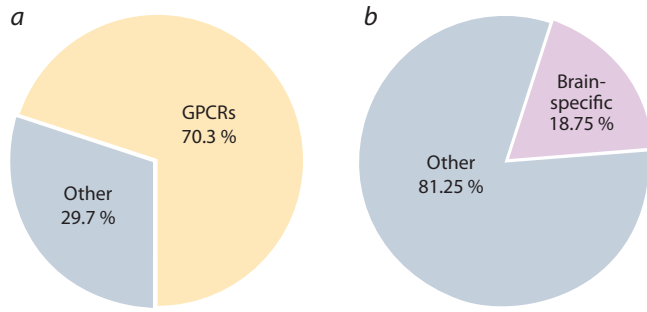


Fig. 1. Functional characteristics of human genes encoding cell surface receptors and involved in appetite regulation (genes from the *Receptors_64* set).

a, The proportion of genes encoding GPCRs; *b*, the proportion of genes that have brain-specific expression pattern (tissue-specific genes were identified using the TSEA tool).

Analysis of evolutionary characteristics

Phylostratigraphic age of genes (PAI-based analysis). At the first step, we have analyzed the distribution of PAI values for all human protein-coding genes (the *allCDS_19,556* set). PAI values were found to be unevenly distributed (Fig. 2, *a*). One third of all genes (33 %) had a PAI equal to zero (cellular organisms, the root of the phylogenetic tree), and the proportions of genes that had PAI values equal to 5 (the stage of vertebrate divergence) and 6 (the stage of euteleostomi divergence) were 17 % and 14 %, respectively. When we examined the distribution of PAI values for a set of human genes encoding cell surface receptors and involved in appetite regulation (the *Receptors_64* set, Supplementary Material 4), we found that 31 genes out of 64 (i. e. 48 %) had PAI values equal to 5 (the stage of vertebrate divergence) (see Fig. 2, *a*). And this number was significantly higher ($p < 0.001$) than the expected number (10.898) calculated based on the distribution obtained for the *allCDS_19,556* gene set (see Fig. 2, *a*, Supplementary Material 5).

As noted above, a large proportion of genes from the *Receptors_64* set (45 genes out of 64) encode GPCRs (see Fig. 1, *a*). To find out whether the evolutionary features of genes from the *Receptors_64* set are caused by the features of genes from the GPCRs superfamily, we analysed the distribution of PAI values for a set of 389 human genes encoding GPCRs represented in the GPCRdb database (https://gpcrdb.org) (*allGPCR_389* set). This distribution was also found to be different from the distribution obtained for all human protein-coding genes (see Fig. 2, *b*). The number of genes in the *allGPCR_389* set that had PAI values equal to 5 (the stage of vertebrate divergence) was 39 % (150 genes out of 389) and it was significantly higher than the expected number calculated based on the proportion of this group of genes in the *allCDS_19,566* set (Supplementary Material 6).

Next, we compared the distribution of PAI values for 45 genes encoding GPCRs and regulating appetite (the *appGPCR_45* set) with the distribution for the *allGPCR_389* set (see Fig. 2, *c*). In the group of genes from the *appGPCR_45* set, 28 genes were found to have a PAI equal to 5 (the stage of vertebrate divergence) (i. e. 64 %), which was significantly higher than the expected number (17.35) calculated based on the proportion of this group of genes in the *allGPCR_389* set (Supplementary Material 7).

As mentioned earlier, 19 receptor genes controlling appetite did not belong to the G-protein-coupled receptor superfamily (the *app_not_GPCR_19* set). When the distribution over PAI values for this group of genes was examined, it was also found to differ from the distribution over PAI values for all human protein-coding genes (see Fig. 2, *d*). However, in this case, a significant ($p < 0.05$) excess over the expected number of genes with a PAI equal to 6 (the stage of euteleostomi divergence) was observed. The observed number in the *app_not_GPCR_19* set was six genes out of 19 (32 %), whereas in the *allCDS_19,566* set, PAI value equal to 6 was detected for 2,769 genes (14 %). Thus, the expected number of genes with PAI equal to 6 in the *app_not_GPCR_19* set was 2.69 (Supplementary Material 8).

Evolutionary variability of genes (DI-based analysis). The analysis of the distribution of genes from the *Receptors_64* set according to DI values (Fig. 3, *a*, Supplementary Material 9) showed that 47 % of genes (30 out of 64) had DI < 0.2, most

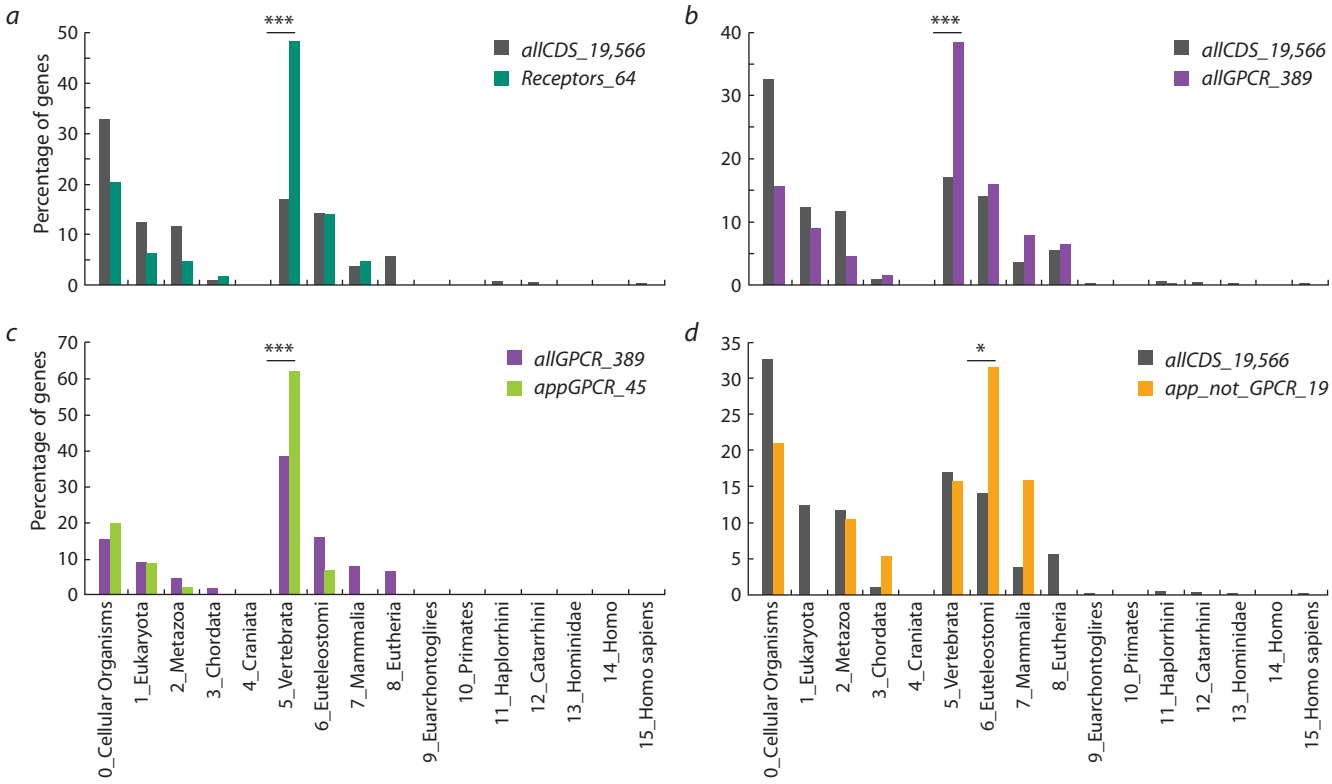


Fig. 2. Distributions by PAI values obtained for the sets of human protein-coding genes presented in Table 1.
a: all human protein-coding genes (*allCDS_19,566*) as a control set and the human appetite-regulating genes encoding receptors (*Receptors_64*); *b*: all human protein-coding genes (*allCDS_19,566*) as a control set and genes encoding GPCRs (*allGPCR_389*); *c*: genes encoding GPCRs as a control set (*allGPCR_389*) and genes encoding GPCRs controling appetite (*appGPCR_45*); *d*: all human protein-coding genes (*allCDS_19,566*) as a control set and genes controlling appetite but not belonging to the GPCRs superfamily (*app_not_GPCR_19*).

PAI values were calculated at a threshold of 0.5 for the level of similarity between the DNA sequences of the orthologous genes. Asterisks indicate differences between the number of genes with a PAI equal to 5 (the stage of vertebrate divergence) (*a–c*) or a PAI equal to 6 (the stage of euteleostomi divergence) (*d*) and their expected numbers calculated based on the distributions in the control sets. *** $p < 0.001$, * $p < 0.05$. See Supplementary Materials 5–8.

genes (63 out of 64, i. e. ~98 %) had DI < 1, and only one gene (*QRFPR*) had DI > 1, indicating that most of the genes are subjected to stabilizing selection.

Comparison of the distribution of genes from the *Receptors_64* set by DI values with the distribution obtained for all human protein-coding genes (*allCDS_19,566* set) showed that the *Receptors_64* set is characterised by an increased content of genes with low DI values (see Fig. 3, *a*). The majority of genes from the *Receptors_64* set (61 genes out of 64, i. e. 95 %) had DI < 0.6. And this number was significantly ($p < 0.01$) higher than the expected number (51.95) calculated using the distribution obtained for all human protein-coding genes (see Fig. 3, *a*, Supplementary Material 10).

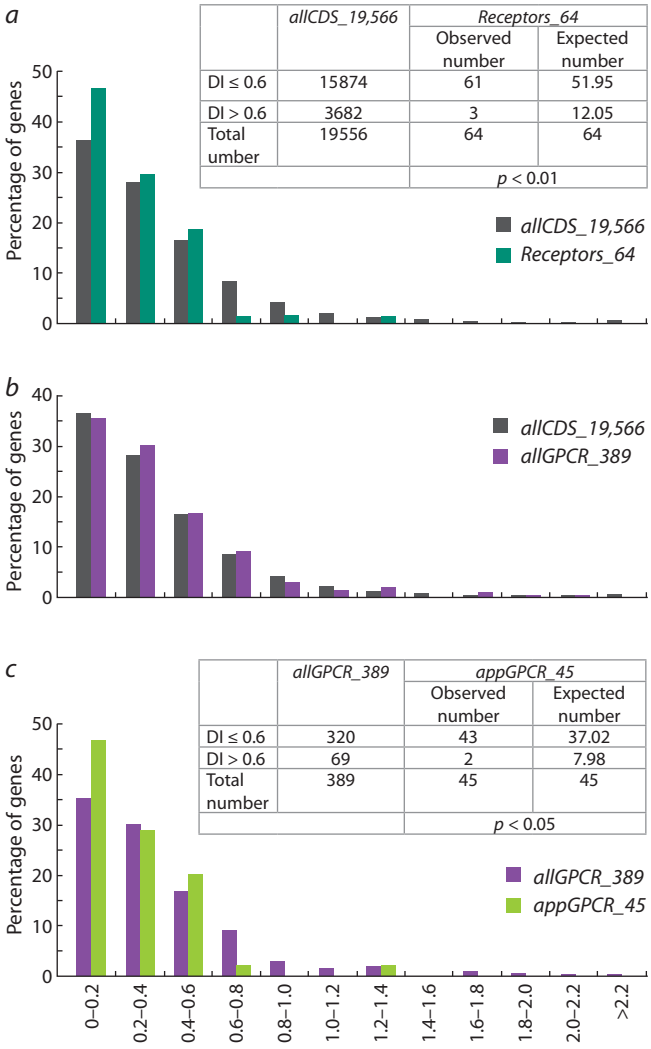
When comparing the distribution over DI values for a set of all receptors from the GPCRs superfamily (*allGPCR_389*) with the distribution for all protein-coding genes (*allCDS_19,566*), no significant differences were found (see Fig. 3, *b*).

Comparison of the distribution over DI values for the *appGPCR_45* set with the distribution for all receptors from the GPCRs superfamily (*allGPCR_389*) showed that the number of genes with low DI (DI ≤ 0.6) in the *appGPCR_45* set (42 genes) was significantly ($p < 0.05$) higher than the expected number of genes (37.018) calculated from the DI distribution for all genes encoding GPCRs (see Fig. 3, *c*, Supplementary Material 11).

As indicated above, approximately one-fifth (18.75 %) of genes from the *Receptors_64* set are brain-specific. We have determined the content of genes that had brain-specific expression patterns in two subgroups of genes: (1) a subgroup of genes with low DI (DI ≤ 0.2); (2) a subgroup including all other genes (they had DI values between 0.2 and 1.3). It turned out that the number of brain-specific genes in these subgroups differs significantly from the expected number calculated based on random distribution: in the subgroup of genes with low DI, the content of brain-specific genes was increased (Fig. 4, Supplementary Material 12).

Discussion

Cell surface receptor genes constitute a substantial proportion (more than a thousand genes) of the human genome (Bausch-Fluck et al., 2018). The interest in the study of cell surface receptors is due to their important role in the cell. These transmembrane proteins interact with various molecules (ligands) located in the extracellular space and activate signal transduction pathways in the cell (Bausch-Fluck et al., 2018; Yang et al., 2021). A lot of substances and biochemical compounds (in particular, drugs) that affect the activity of cell surface



	allCDS_19,566	Receptors_64	
		Observed number	Expected number
DI ≤ 0.6	15874	61	51.95
DI > 0.6	3682	3	12.05
Total umber	19556	64	64
		$p < 0.01$	

	allGPCR_389	appGPCR_45	
		Observed number	Expected number
DI ≤ 0.6	320	43	37.02
DI > 0.6	69	2	7.98
Total number	389	45	45
		$p < 0.05$	

Fig. 3. Distributions of genes from the sets presented in Table 1 according to the DI index.

a, All human protein-coding genes (*allCDS_19,566*) as a control set and and the human appetite-regulating genes encoding receptors (*Receptors_64*). The observed and expected total number of genes with DI ≤ 0.6 and DI > 0.6 are presented in the table above the graph; the calculation of the expected number is given in Supplementary Material 10. *b*, All human protein-coding genes (*allCDS_19,566*) as a control set and genes encoding GPCRs (*allGPCR_389*). *c*, Genes encoding GPCRs (*allGPCR_389*) as a control set and genes encoding GPCRs controling appetite (*appGPCR_45*). The observed and expected total number of genes with DI ≤ 0.6 and DI > 0.6 are presented in the table above the graph; the calculation of the expected number is given in Supplementary Material 11.

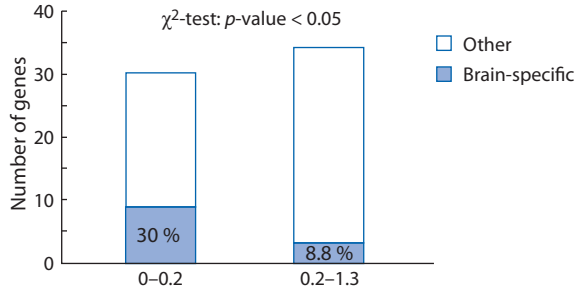


Fig. 4. Distribution of genes from the *Receptors_64* set by DI values.

Shown are the proportions of genes that have brain-specific expression patterns according to the TSEA tool. The observed number of brain-specific genes differ from the expected number, * $p < 0.05$. (The numbers of genes in four subgroups are given in Supplementary Material 12.)

receptors (so-called agonists and antagonists) are known. Therefore, cell surface receptors are also of great interest from a biomedical point of view – for example, these proteins are targets for 66 % of drugs registered in the DrugBank database (Bausch-Fluck et al., 2018).

This paper presents a set of 64 human genes encoding cell surface receptors, the orthologs of which are involved in food intake regulation in model organisms. The data are highly reliable, because this gene set was created on the basis of manual analysis of scientific publications. Finding such an impressive number of receptor genes involved in appetite regulation fits well with the idea of the complex nature of food motivation. As mentioned above, appetite can satisfy both the basic needs of the organism for food (homeostatic appetite, which provides compensation of energy expenditure) and the needs for sensations associated with food (non-homeostatic appetite, aimed at obtaining positive emotions) (Johnson, 2013; Rebello, Greenway, 2016; Ahn et al., 2022). It is also known that food motivation can be adjusted depending on the life situation or psycho-emotional state of an individual (fright, depression, boredom, chronic stress, for animals – threat from predators, territory protection, mating behavior, etc.) (Lindén et al., 1987; Braden et al., 2018, 2023; Hadjieconomou et al., 2020; Siegal et al., 2022). Such correction of food motivation is performed because the brain processes information received from the sensory organs and integrates it with signals about the state of various physiological systems of the body (Tomé et al., 2009; Holtmann, Talley, 2014; Spetter et al., 2014; Tremblay, Bellisle, 2015). And this process involves nerve cells with diverse specialization expressing a wide range of receptors on their surface (Yeo, Heisler, 2012; Heisler, Lam, 2017).

Examination of the distributions of genes by PAI values revealed that: (1) the *Receptors_64* set has a significantly increased content of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence) than all protein-coding genes; (2) the subset of genes that encode GPCRs and are involved in appetite regulation (*appGPCR_45*) also contains an increased number of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence) than what would be expected based on the distribution of PAI values for all genes encoding GPCRs.

Thus, we found that the gene set composed of genes encoding cell surface receptors controlling appetite contains an increased number of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence). Apparently, the synchronised evolution of such a large group of genes (31 genes have PAIs equal to five) is associated with the formation of the brain as a separate organ in the first vertebrates (Sarnat, Netsky, 2002). It is noteworthy that almost all of these genes with PAI equal to five (28 out of 31) encode GPCRs, which agrees well with the fact that receptors of this superfamily are involved in processing signals from sensory organs, as well as signals transmitted by hormones and neuromediators (Pandy-Szekeres et al., 2023). Thus, the group of

genes encoding GPCRs with a PAI of 5 (the stage of vertebrate divergence) includes, in particular, genes encoding receptors for neuropeptide Y (*NPY1R*, *NPY2R*, *NPY4R*, *NPY5R*) and alpha-melanocyte stimulating hormone (*MC3R* and *MC4R*). Neuropeptide Y and alpha-melanocyte stimulating hormone are signalling molecules secreted by neurons of the arcuate nucleus of the hypothalamus, a brain structure that acts as a central regulator of feeding behaviour (Yeo, Heisler, 2012; Heisler, Lam, 2017).

Another peculiarity was revealed for a subset of genes involved in appetite regulation but not encoding GPCRs (*app_not_GPCR_19*): it contains an increased number of genes with PAI equal to 6 (the stage of euteleostomi divergence). Notably, four genes from this group encode receptors involved in the regulation of immunity. These are *GHR* and *LEPR* encoding proteins from the type I cytokine receptor family and *TLR2* and *TLR4* encoding proteins from the Toll-like receptor family.

The PAI-based analysis has shown that the so-called "ancient" genes (i. e., genes with PAI equal to 0 (cellular organisms, the root of the taxonomic tree)) are also involved in the regulation of food intake. This group includes, for example, genes encoding (1) the insulin receptor (*INSR*), which, in particular, regulates secretion of neuropeptide Y and alpha-melanocyte stimulating hormone by neurons of the arcuate nucleus of the hypothalamus (Leibowitz, Wortley, 2004), and (2) *NTRK2*, the receptor for BDNF (brain-derived neurotrophic factor), which mediates the anorexigenic effects of BDNF produced in the paraventricular nucleus of the hypothalamus (An et al., 2015; Chu et al., 2023). Both genes are expressed in different tissues and organs (Escandón et al., 1994; Federici et al., 1997), indicating that at early stages of evolution, ancestral forms of *INSR* and *NTRK2* could be involved in the regulation of various biological processes and joined the system of genes regulating food intake at the evolutionary stage corresponding to the formation of specialised brain structures.

When considering PAI values, a group of relatively "young" genes was identified (PAI values of 6 and 7, the stages of euteleostomi and mammalia divergence). Five genes from this group encode receptors relevant to the immune system: these are the four genes mentioned earlier (*GHR*, *LEPR*, *TLR2*, and *TLR4*), as well as *IL1R1*. The detection of these genes in a group of relatively "young" genes is in good agreement with the known data on the adaptive immunity system having begun to form relatively recently in the course of evolution (Ward, Rosenthal, 2014).

When examining the distribution of genes from the *Receptors_64* set by DI values, a significant enrichment of this group with genes subjected to stabilizing selection was revealed. It turned out that the subgroup of appetite-regulating genes encoding GPCRs (*appGPCR_45*) also contained an increased number of genes with low DI values.

Eight genes had the lowest DI values (DI < 0.05): *GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1* (see Supplementary Material 9). Moreover, seven of these eight genes (except *ADIPOR1*) encode proteins from the GPCRs superfamily.

An extremely low DI value (<0.005) was found for the *GPR26* gene. *GPR26* encodes a receptor from the GPCRs superfamily, the endogenous ligand of which has not yet been identified. Targeted inactivation of *GPR26* in mice causes hyperphagia leading to early onset of diet-induced obesity (Chen et al., 2012). In addition, according to behavioural tests, *Gpr26*-deficient mice display increased anxiety- and depression-like behavior, and prefer ethanol to a greater extent than mice with normal genotype (Zhang et al., 2011). According to the TSEA tool, *GPR26* has brain-specific expression pattern. In humans, *GPR26* is expressed in the amygdala, hippocampus, and thalamus (Jones et al., 2007). The function of the *GPR26* gene is evolutionarily conserved. In *C. elegans*, the *Y5H2B* gene with similarity to *GPR26* was found. Ashrafi K. and co-workers used RNA-mediated interference to disrupt the expression of genes and found that *Y5H2B* increases fat content (Ashrafi et al., 2003). The functions of other genes that had extremely low DI values (<0.05) are described in Supplementary Material 13.

Only one gene (*QRFPR*) among the genes from the *Receptors_64* set had a DI > 1 (see Supplementary Material 9), indicating that this gene is probably subjected to driving selection. *QRFPR* encodes the receptor for the orexigenic neuropeptide QRFP (pyroglutamylated RFamide peptide) (Cook et al., 2022). According to EntrezGene and UniProt databases, human *QRFPR* is expressed in different parts of the brain and in peripheral tissues (heart, kidney, stomach, testes, and thyroid gland). The mouse, rat, and hamster genomes are known to contain at least two genes encoding the neuropeptide receptor QRFP (Cook et al., 2022). No data like that are available for the human genome; however, it can be assumed that human *QRFPR* is not subjected to stabilizing selection, since the human genome also contains more than one gene encoding proteins with QRFPR-like activity.

We also found that the group of genes with low DI, i. e. most likely to be subject to stabilizing selection, is enriched in genes that have brain-specific expression patterns. This result agrees well with the finding made by G. Dumas et al. who examined the set of almost all human protein-coding genes (N = 11,667) and revealed that genes encoding brain-related proteins are among the most strongly conserved protein-coding genes in the human genome (Dumas et al., 2021). Among the genes that have low DI and brain-specific expression pattern, the previously mentioned *GPR26* gene was found. Due to the fact that this gene has an extremely low DI and its endogenous ligand is still unknown (Chen et al., 2012), *GPR26* seems to be an extremely interesting object for further theoretical and experimental studies.

Conclusion

In this paper, we analyzed the distributions of PAI and DI values for a group of human cell surface receptor genes, the orthologues of which were involved in appetite regulation in model organisms. It was found that the gene set under consideration contains an increased number of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence), which is apparently associated with the formation of the brain as a separate organ in the first vertebrates. A significant enrichment of this group of genes in genes with low DI values was also revealed, indicating a significant susceptibility of these genes to stabilizing selection. At the same time, the group of genes with low DI is enriched with genes that have brain-specific expression pattern. The

characteristic features of the cell surface receptor genes distribution according to the evolutionary indices PAI and DI revealed in this study are a starting point for further analyses of the evolutionary characteristics of the entire gene network controlling appetite.

References

Ahn B.H., Kim M., Kim S.Y. Brain circuits for promoting homeostatic and non-homeostatic appetites. *Exp. Mol. Med.* 2022;54(4):349-357. DOI 10.1038/s12276-022-00758-4

An J.J., Liao G.Y., Kinney C.E., Sahibzada N., Xu B. Discrete BDNF neurons in the paraventricular hypothalamus control feeding and energy expenditure. *Cell Metab.* 2015;22(1):175-188. DOI 10.1016/j.cmet.2015.05.008

Ashrafi K., Chang F.Y., Watts J.L., Fraser A.G., Kamath R.S., Ahringer J., Ruvkun G. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature.* 2003;421(6920):268-272. DOI 10.1038/nature01279

Bausch-Fluck D., Goldmann U., Müller S., van Oostrum M., Müller M., Schubert O.T., Wollscheid B. The in silico human surfaceome. *Proc. Natl. Acad. Sci. USA.* 2018;115(46):E10988-E10997. DOI 10.1073/pnas.1808790115

Bjarnadóttir T.K., Gloriam D.E., Hellstrand S.H., Kristiansson H., Fredriksson R., Schiöth H.B. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics.* 2006;88(3):263-273. DOI 10.1016/j.ygeno.2006.04.001

Braden A., Musher-Eizenman D., Watford T., Emley E. Eating when depressed, anxious, bored, or happy: are emotional eating types associated with unique psychological and physical health correlates? *Appetite*. 2018;125:410-417. DOI 10.1016/j.appet.2018.02.022

Braden A., Barnhart W.R., Kalantzis M., Redondo R., Dauber A., Anderson L., Tilstra-Ferrell E.L. Eating when depressed, anxious, bored, or happy: an examination in treatment-seeking adults with overweight/obesity. *Appetite*. 2023;184:106510. DOI 10.1016/j.appet.2023.106510

Chen D., Liu X., Zhang W., Shi Y. Targeted inactivation of GPR26 leads to hyperphagia and adiposity by activating AMPK in the hypothalamus. *PLoS One.* 2012;7(7):e40764. DOI 10.1371/journal.pone.0040764

Chu P., Guo W., You H., Lu B. Regulation of satiety by *Bdnf-e2*-expressing neurons through TrkB activation in ventromedial hypothalamus. *Biomolecules.* 2023;13(5):822. DOI 10.3390/biom13050822

Cook C., Nunn N., Worth A.A., Bechtold D.A., Suter T., Gackeheimer S., Foltz L., Emmerson P.J., Statnick M.A., Luckman S.M. The hypothalamic RFamide, QRFP, increases feeding and locomotor activity: the role of Gpr103 and orexin receptors. *PLoS One.* 2022;17(10):e0275604. DOI 10.1371/journal.pone.0275604

Dumas G., Malesys S., Bourgeron T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Res.* 2021;31(3):484-496. DOI 10.1101/gr.262113.120

Escandón E., Soppet D., Rosenthal A., Mendoza-Ramírez J.L., Szönyi E., Burton L.E., Henderson C.E., Parada L.F., Nikolics K. Regulation of neurotrophin receptor expression during embryonic and postnatal development. *J. Neurosci.* 1994;14(4):2054-2068. DOI 10.1523/JNEUROSCI.14-04-02054.1994

Federici M., Porzio O., Zucaro L., Fusco A., Borboni P., Lauro D., Sesti G. Distribution of insulin/insulin-like growth factor-I hybrid receptors in human tissues. *Mol. Cell. Endocrinol.* 1997;129(2):121-126. DOI 10.1016/s0303-7207(97)04050-1

Fichter M.M., Quadflieg N. Mortality in eating disorders – results of a large prospective clinical longitudinal study. *Int. J. Eat. Disord.* 2016;49(4):391-401. DOI 10.1002/eat.22501

Grossberg A.J., Scarlett J.M., Marks D.L. Hypothalamic mechanisms in cachexia. *Physiol. Behav.* 2010;100(5):478-489. DOI 10.1016/j.physbeh.2010.03.011

GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660. DOI 10.1126/science.1262110

Hadjieconomou D., King G., Gaspar P., Mineo A., Blackie L., Ameku T., Studd C., de Mendoza A., Diao F., White B.H., Brown A.E.X., Plaçais P.Y., Préat T., Miguel-Aliaga I. Enteric neurons increase maternal food intake during reproduction. *Nature.* 2020;587(7834):455-459. DOI 10.1038/s41586-020-2866-8

Heisler L.K., Lam D.D. An appetite for life: brain regulation of hunger and satiety. *Curr. Opin. Pharmacol.* 2017;37:100-106. DOI 10.1016/j.coph.2017.09.002

Holtmann G., Talley N.J. The stomach-brain axis. *Best Pract. Res. Clin. Gastroenterol.* 2014;28(6):967-979. DOI 10.1016/j.bpg.2014.10.001

Ignatieva E.V., Afonnikov D.A., Rogaev E.I., Kolchanov N.A. Human genes controlling feeding behavior or body mass and their functional and genomic characteristics: a review. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2014;18(4/2):867-875 (in Russian)

Ignatieva E.V., Afonnikov D.A., Saik O.V., Rogaev E.I., Kolchanov N.A. A compendium of human genes regulating feeding behavior and body weight, its functional characterization and identification of GWAS genes involved in brain-specific PPI network. *BMC Genet.* 2016;17(Suppl.3):158. DOI 10.1186/s12863-016-0466-2

Johnson A.W. Eating beyond metabolic need: how environmental cues influence feeding behavior. *Trends Neurosci.* 2013;36(2):101-109. DOI 10.1016/j.tins.2013.01.002

Jones P.G., Nawoschik S.P., Sreekumar K., Uveges A.J., Tseng E., Zhang L., Johnson J., He L., Paulsen J.E., Bates B., Pausch M.H. Tissue distribution and functional analyses of the constitutively active orphan G protein coupled receptors, GPR26 and GPR78. *Biochim. Biophys. Acta.* 2007;1770(6):890-901. DOI 10.1016/j.bbagen.2007.01.013

Kaidar-Person O., Bar-Sela G., Person B. The two major epidemics of the twenty-first century: obesity and cancer. *Obes. Surg.* 2011;21(11):1792-1797. DOI 10.1007/s11695-011-0490-2

Katritch V., Cherezov V., Stevens R.C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* 2013;53:531-556. DOI 10.1146/annurev-pharmtox-032112-135923

Leibowitz S.F., Wortley K.E. Hypothalamic control of energy balance: different peptides, different functions. *Peptides.* 2004;25(3):473-504. DOI 10.1016/j.peptides.2004.02.006

Lindén A., Hansen S., Bednar I., Forsberg G., Södersten P., Uvnäs-Moberg K. Sexual activity increases plasma concentrations of cholecystokinin octapeptide and offsets hunger in male rats. *J. Endocrinol.* 1987;115(1):91-95. DOI 10.1677/joe.0.1150091

Maniam J., Morris M.J. The link between stress and feeding behaviour. *Neuropharmacology.* 2012;63(1):97-110. DOI 10.1016/j.neuropharm.2012.04.017

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2021;25(1):46-56. DOI 10.18699/VJ21.006

New D.C., Wong J.T. The evidence for G-protein-coupled receptors and heterotrimeric G proteins in protozoa and ancestral metazoa. *Biol. Signals Recept.* 1998;7(2):98-108. DOI 10.1159/000014535

Olszewski P.K., Cedernaes J., Olsson F., Levine A.S., Schiöth H.B. Analysis of the network of feeding neuroregulators using the Allen Brain Atlas. *Neurosci. Biobehav. Rev*. 2008;32(5):945-956. DOI 10.1016/j.neubiorev.2008.01.007

Pandy-Szekeres G., Caroli J., Mamyrbekov A., Kermani A.A., Keseru G.M., Kooistra A.J., Gloriam D.E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Res.* 2023;51(D1):D395-D402. DOI 10.1093/nar/gkac1013

Rebello C.J., Greenway F.L. Reward-induced eating: therapeutic approaches to addressing food cravings. *Adv. Ther.* 2016;33(11):1853-1866. DOI 10.1007/s12325-016-0414-6

Sarnat H.B., Netsky M.G. When does a ganglion become a brain? Evolutionary origin of the central nervous system. *Semin. Pediatr. Neurol.* 2002;9(4):240-253. DOI 10.1053/spen.2002.32502

Siegal E., Hooker S.K., Isojunno S., Miller P.J.O. Beaked whales and state-dependent decision-making: how does body condition affect the trade-off between foraging and predator avoidance? *Proc. Biol. Sci.* 2022;289(1967):20212539. DOI 10.1098/rspb.2021.2539

Spetter M.S., de Graaf C., Mars M., Viergever M.A., Smeets P.A. The sum of its parts – effects of gastric distention, nutrient content and sensory stimulation on brain activation. *PLoS One.* 2014;9(3): e90872. DOI 10.1371/journal.pone.0090872

Tomé D., Schwarz J., Darcel N., Fromentin G. Protein, amino acids, vagus nerve signaling, and the brain. *Am. J. Clin. Nutr.* 2009;90(3): 838S-843S. DOI 10.3945/ajcn.2009.27462W

Tremblay A., Bellisle F. Nutrients, satiety, and control of energy intake. *Appl. Physiol. Nutr. Metab.* 2015;40(10):971-979. DOI 10.1139/apnm-2014-0549

Ward A.E., Rosenthal B.M. Evolutionary responses of innate immunity to adaptive immunity. *Infect. Genet. Evol.* 2014;21:492-496. DOI 10.1016/j.meegid.2013.12.021

Wells A., Kopp N., Xu X., O'Brien D.R., Yang W., Nehorai A., Adair-Kirk T.L., Kopan R., Dougherty J.D. The anatomical distribution of genetic associations. *Nucleic Acids Res.* 2015;43(22):10804-10820. DOI 10.1093/nar/gkv1262

Yang D., Zhou Q., Labroska V., Qin S., Darbalaei S., Wu Y., Yuliantie E., Xie L., Tao H., Cheng J., Liu Q., Zhao S., Shui W., Jiang Y., Wang M.W. G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct. Target. Ther.* 2021;6(1):7. DOI 10.1038/s41392-020-00435-w

Yeo G.S., Heisler L.K. Unraveling the brain regulation of appetite: lessons from genetics. *Nat. Neurosci.* 2012;15(10):1343-1349. DOI 10.1038/nn.3211

Zhang L.L., Wang J.J., Liu Y., Lu X.B., Kuang Y., Wan Y.H., Chen Y., Yan H.M., Fei J., Wang Z.G. GPR26-deficient mice display increased anxiety- and depression-like behaviors accompanied by reduced phosphorylated cyclic AMP responsive element-binding protein level in central amygdala. *Neuroscience.* 2011;196:203-214. DOI 10.1016/j.neuroscience.2011.08.069

ORCID ID

E.V. Ignatieva orcid.org/0000-0002-8588-6511
S.A. Lashin orcid.org/0000-0003-3138-381X
Z.S. Mustafin orcid.org/0000-0003-2724-4497
N.A. Kolchanov orcid.org/0000-0001-6800-8787

Acknowledgements. The work was supported by the publicly funded project No. FWNR-2022-0020 of the Federal Research Center ICG SB RAS.

Conflict of interest. The authors declare no conflict of interest.

Received August 11, 2023. Revised September 5, 2023. Accepted September 7, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):839-850
DOI 10.18699/VJGB-23-97

Original Russian text https://vavilovj-icg.ru/

On the space of SARS-CoV-2 genetic sequence variants

A.Yu. Palyanov[1, 2, 3] ✉, N.V. Palyanova[2]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Virology, Federal Research Center of Fundamental and Translational Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ palyanov@iis.nsk.su

Abstract. The coronavirus pandemic caused by the SARS-CoV-2 virus, which humanity resisted using the latest advances in science, left behind, among other things, extensive genetic data. Every day since the end of 2019, samples of the virus genomes have been collected around the world, which makes it possible to trace its evolution in detail from its emergence to the present. The accumulated statistics of testing results showed that the number of confirmed cases of SARS-CoV-2 infection was at least 767.5 million (9.5 % of the current world population, excluding asymptomatic people), and the number of sequenced virus genomes is more than 15.7 million (which is over 2 % of the total number of infected people). These new data potentially contain information about the mechanisms of the variability and spread of the virus, its interaction with the human immune system, the main parameters characterizing the mechanisms of the development of a pandemic, and much more. In this article, we analyze the space of possible variants of SARS-CoV-2 genetic sequences both from a mathematical point of view and taking into account the biological limitations inherent in this system, known both from general biological knowledge and from the consideration of the characteristics of this particular virus. We have developed software capable of loading and analyzing SARS-CoV-2 nucleotide sequences in FASTA format, determining the 5′ and 3′ UTR positions, the number and location of unidentified nucleotides ("N"), performing alignment with the reference sequence by calling the program designed for this, determining mutations, deletions and insertions, as well as calculating various characteristics of virus genomes with a given time step (days, weeks, months, etc.). The data obtained indicate that, despite the apparent mathematical diversity of possible options for changing the virus over time, the corridor of the evolutionary trajectory that the coronavirus has passed through seems to be quite narrow. Thus it can be assumed that it is determined to some extent, which allows us to hope for a possibility of modeling the evolution of the coronavirus.
Key words: coronavirus; SARS-CoV-2; genome; space of variants; evolution; variability.

For citation: Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):839-850. DOI 10.18699/VJGB-23-97

О пространстве вариантов генетических последовательностей SARS-CoV-2

А.Ю. Пальянов[1, 2, 3] ✉, Н.В. Пальянова[2]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт вирусологии, Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ palyanov@iis.nsk.su

Аннотация. Пандемия коронавирусной инфекции, вызванная вирусом SARS-CoV-2, которой человечество противостояло с использованием новейших достижений науки, оставила после себя в том числе обширные генетические данные. Ежедневно начиная с конца 2019 г. в мире собирались образцы геномов вируса, что предоставляет возможность детально проследить его эволюцию с момента возникновения до настоящего времени. Накопленная статистика результатов экспресс-тестирования показала, что число подтвержденных случаев заражения SARS-CoV-2 составило не менее 767.5 млн (9.5 % нынешнего населения Земли без учета бессимптомников), а число секвенированных геномов вируса – более 15.7 млн (что составляет чуть более 2 % от общего числа заразившихся). Эти новые данные потенциально несут в себе информацию о механизмах изменчивости и распространения вируса, его взаимодействия с иммунной системой человека, об основных параметрах, характеризующих механизмы развития пандемии, и многое другое. В этой статье мы анализируем пространство возможных вариантов генетических последовательностей SARS-CoV-2 как с математической точки зрения, так и с учетом биологических ограничений, присущих этой системе (основанных на общебиологических знаниях и учитывающих особенности данного конкретного вируса). Для этого мы разработали

программное обеспечение, способное загружать и анализировать нуклеотидные последовательности SARS-CoV-2 в формате FASTA, определять позиции 5′ и 3′ UTR, число и расположение неидентифицированных нуклеотидов (“N”), осуществлять выравнивание относительно референсной последовательности посредством вызова предназначенных для этого программ, определять мутации, делеции и вставки, а также рассчитывать различные характеристики геномов вирусов с заданным шагом по времени (дни, недели, месяцы и т. д.). Полученные данные свидетельствуют о том, что, несмотря на кажущееся математическое многообразие возможных вариантов изменения вируса во времени, коридор эволюционной траектории, которым прошел коронавирус, представляется достаточно узким. Это дает основание полагать, что он в некоторой степени детерминирован, что позволяет надеяться на возможность моделирования эволюции коронавируса.
Ключевые слова: коронавирус; SARS-CoV-2; геном; пространство вариантов; эволюция; изменчивость.

Introduction

The possibility of computational modeling of the evolution, life cycle and reproduction of the simplest biological organism down to the gene level would be a scientific breakthrough, but it is still far beyond the capabilities of modern supercomputers. The process of natural selection of the fittest individuals takes into account a huge number of factors in both the external and internal environment. The characteristics of an organism are realized through sets of protein characteristics and features, and the impact of changes in each protein on an organism's fitness is quite difficult to assess due to the need to take into account all the resulting changes in the interactions of a protein with all environmental factors and other proteins, the number of which is very significant.

Usually, in computer models of evolving objects, changes to the genome of descendants are not carried out directly (by reproducing molecular mechanisms), but are only simulated by describing algorithms for making changes to a copy of the genome of ancestors. However, the mechanisms of introducing mutations and horizontal gene transfer themselves are subjects of evolution, and among possible changes that do not lead to the death or sterility of an individual, there are also those that affect the speed and accuracy of genome replication. Due to this, intraspecific competition arises, as a result of which, for example, in the case of SARS-CoV-2, from the moment of its appearance to the present time, the duration of the incubation period, directly related to the rate of virus replication, is constantly decreasing (Malone et al., 2022).

In comparison with cellular life forms, viruses are substantially simpler and thus are much more convenient for investigation and computational modeling of their evolution, especially taking into account their significantly smaller genomes and, at the same time, still quite wide range of interactions with the external environment and host organism. Before the appearance of fast genome sequencing technologies, evolution of viruses could only be considered within the framework of “parasite–host” models, which described statistical, but not molecular features of their interaction. Since the beginning of the SARS-CoV-2 pandemic, the number of confirmed cases of this infection has been at least 767.5 million (9.5 % of the current world population, excluding asymptomatic people) (Palyanova et al., 2022). During this period, the global scientific community has obtained more than 15.7 million variants of the genomes of this coronavirus (including the date of sampling and the geographical location of the place where it was collected), providing unprecedentedly extensive data on its evolution, in such quantities that were not available for any other virus before.

Based on these data, the dynamics of spread and change of the virus can be calculated not only in physical space and time, but also in the multidimensional space of possible viable variants of viral genomes with a metric determined by the minimum number of single changes (mutation, deletion or insertion) required to transform one genome into another (known as the “Levenshtein distance” or “edit distance”). The virus changes over time, including the response to vaccination and the formation of immunity in people who have recovered from the disease. This means that both the genome of the virus and its “phenotypic” manifestations change when interacting with the carrier's body, i. e. two parallel processes occur simultaneously – both a change (spread) of a set (cloud) of points representing the virus population (at one time or another) in the space of possible RNA sequences, and a change in the very landscape of this multidimensional surface of the “fitness function” of the virus. Each point in the space of possible states corresponds to a specific nucleotide sequence, more or less different from the original reference genome (from which it all began at the end of 2019 (Wu et al., 2020)) by a certain number of changes – mutations, deletions and insertions.

Transitions can and should exist between pairs of points (in the space of viral RNA sequence variants), each of which corresponds to a viable sequence, if one of them has resulted from the other via changes that have occurred within the virus from the moment it enters the host's body until the appearance of the next generation of virions (usually many more than one cycle of replication of the virus genome takes place before this). Most of the possible changes that occur during replication (each copy of the viral sequence has its own set) will lead to the appearance of a non-viable variant (especially deletions or insertions the length of which is not a multiple of three – i. e. those that will lead to a reading frame shift during translation). However, some changes can leave the fitness of the virus at the same level or even increase it – for example, by raising the rate of synthesis of new viral particles or increasing their number per time unit (which will

give them advantage over other variants located in the body at the same time, i. e. intraspecific competition arises). The fitness function of some viral sequence can be thought of as the number of its copies existing in the human population at a given time (with or without normalization to the total number of virus copies).

Thus, the landscape of the "surface" of the (multidimensional) fitness function is formed, which may have more or less extensive "valleys" corresponding to many similar sequences (appearing as a result of small changes in the variant that first fell into this valley), surrounded by "mountains". There are "mountain ridges" or "plateaus" (all points of which correspond to non-viable sequences) delimiting "valleys" of viable sequences and "passes" between them. Regions of non-viable sequences correspond to the cases when, for example, a virus cannot make copies of itself due to damage to the gene encoding RNA-dependent RNA polymerase (RdRp), which performs viral RNA replication, or when changes in the structure of the capsid proteins prevents the virus from forming a protein shell, as well as for many other diverse reasons. Also, presumably, there are "valleys" for which none of the sequence variants belonging to them have yet been realized, but which can be reached in the future – for example, due to the emergence of a viable recombinant strain resulting from a combination of the genomes of two not very similar variants of the virus. It is possible that this is how the initial WT strain of SARS-CoV-2 arose.

There are currently two major databases providing online access to SARS-CoV-2 genetic sequences. The largest of them is GISAID (https://gisaid.org) – Global Initiative on Sharing All Influenza Data (started in 2006) (Khare et al., 2021). Since the emergence of SARS-CoV-2 at the end of 2019, it has also become a repository for the accumulation of sequenced variants of this virus obtained by laboratories around the world. In July 2023, there are more than 15.7 million SARS-CoV-2 sequences stored in it. Another database, NCBI SARS-CoV-2 Data Hub (Sayers et al., 2022) (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?VirusLineage_ss=taxid:2697049), contains more than 7.7 million SARS-CoV-2 genome sequences. Such unprecedentedly vast and detailed data have never been available to humanity before, so it is necessary to extract as much useful information and knowledge as possible from their comprehensive analysis. In this work we consider only the first steps on this path, and much remains to be done.

The Nextstrain/Nextclade project (https://clades.nextstrain.org) (Aksamentov et al., 2021), which provides online tools for analysis and visualization of genetic data on various viruses, including SARS-CoV-2, is also of great importance for the scientific community of viral genome researchers. Nextclade's functionality stands out by providing a graphical representation of the genome map of the loaded sequences, showing mutations, deletions, insertions, unidentified nucleotides ("N") and a number of other features of each sequence, including, for example, detection of reassortant (recombinant) variants.

The description of the space of variants of SARS-CoV-2 genetic sequences fundamentally includes (a) those that we can already observe and study thanks to extensive sequencing, (b) variants from the real space of variants that have already been implemented, but have not come to the attention of researchers, and (c) other possible variants that could be realized in the future and are of particular interest, since they are potentially dangerous to humanity and it would be good to be prepared in advance for their possible appearance (rapid tests for their detection, vaccines, etc.).

Let's now consider the most important characteristics of SARS-CoV-2 as a system, the basis of which is self-replication in the host cells, and which may be important in the future when creating its evolutionary simulator. They include the speed of genome replication (600–700 nt/s, the highest among the known speeds of viral RNA polymerases) (Shannon et al., 2020), the time of viral RNA replication ($\frac{3 \cdot 10^4 \text{ nt}}{600 \text{ nt/s}} = 50$ s), the entire virus reproduction time (7–24 hours) (Grebennikov et al., 2021) and the frequency of replication errors occurrence ($1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ per position, per cycle of cell infection, i. e. from the entry of the virus into a cell until the release of new virions out of it) (Amicone et al., 2022). The rate of its evolution is estimated as $8.9 \cdot 10^{-4}$ changes per position per year (Sonnleitner, 2022), which could lead to an average of 93 changes in 3.5 years. This correlates quite well with the fact that one of the variants most distant from the reference sequence (belonging to the "Omicron" variant, obtained on June 20, 2023) has 103 substitutions (the maximum number of mutations among the variants, see the Table). The "Alpha" and the "Beta" variants differ from the reference sequence by more than 30 point mutations and more than 17 deletions. The variants that arose later have more differences. It is also noticeable that during the evolution of the virus the number of deletions increases, reaching 59 in one of the recent branches of "Omicron".

As was already mentioned, the SARS-CoV-2 coronavirus has the fastest RNA polymerase, but it also has one of the lowest (for RNA viruses) rates of mutation occurrence during the replication process, which is necessary due to its rather large genome. This is achieved thanks to the error-correcting exonuclease (nsp14-ExoN), which is found only in viruses with large genomes (coronaviruses and toroviruses) (Campagnola et al., 2022).

Also among important parameters are the minimum infectious dose (the number of virions required for infection), which is about 100 particles (Karimzadeh et al., 2021), the reproductive number (1.8–3.2) (Xu et al., 2021), the number of viral particles carried by a patient during the peak of infection ($(1–100) \cdot 10^9$) and the number of virions contained on average in an infected cell (10^5) (Sender et al., 2021), as well as other epidemiological characteristics. Viral particles are found in many tissues and organs, from the lungs to the brain, but only those present in the respiratory tract or intestines will be released and can be transmitted to subsequent

The most recent representatives of various branches of the phylogenetic tree of coronavirus SARS-CoV-2 (https://nextstrain.org/ncov/open/global/all-time)

Name	Collection date	Accession ID	Pangolin Pango Lineage	Clade, Emerging Lineage	Muta-tions	Gaps, bp	Genome length
hCoV-19/Wuhan/ WIV04/2019 (reference sequences in GISAID)	30.12.2019	EPI_ISL_402124	B	19A	0	0	29 891
Wuhan-Hu-1 (reference sequences in Genbank)	12.2019	NC_045512.2	B	19A	0	0	29 903
hCoV-19/Tunisia/ S-1180/2021	29.10.2021	EPI_ISL_11333927	B.1.1.7	20I (Alpha, V1)	37	19	29 758
hCoV-19/Madagascar/LA2M-112753/2021	16.01.2021	EPI_ISL_7722749	B.1.351.2	20H (Beta, V2)	31	18	29 818
PHL/COVID-74517/2021	01.07.2021	OL629469	B.1.351	20H (Beta, V2)	32	9	29 854
hCoV-19/Brazil/AM-IMTSP-CD24003/2021	10.08.2021	EPI_ISL_14800432	P.1.4	20J (Gamma, V3)	42	9	29 772
LAO/LOMWRU-0461/2021	24.11.2021	OQ028273	P.1	20J (Gamma, V3)	32	18	29 699
hCoV-19/Australia/WA11930/2023	28.02.2023	EPI_ISL_17187319	XBC.1.4	21I (Delta) XBC	77	36	29 308
hCoV-19/Yunnan/ YNCDC-1019/2023	23.05.2023	EPI_ISL_17778593	DY.1	22B (Omicron)	89	59	29 806
hCoV-19/Japan/TKYmbc38047/2023	06.06.2023	EPI_ISL_17941095	XBB.2.3.11	22F (XBB.2.3)	99	56	29 726
hCoV-19/Heilong-jiang/HLJCDC-1665/2023	20.06.2023	EPI_ISL_17850574	XBB.1.5	23A (Omicron) (XBB.1.5)	103	56	29 781

Note. Representatives of some branches (mainly belonging to different variants of "Omicron") are still being found in sequenced specimens of SARS-CoV-2 genomes from recently infected people, and some have ceased to be detected at all ("Alpha", "Beta", "Gamma", "Delta", etc.). The reference sequences in both databases differ only in the length of the poly-A region located at the very end, and in all other positions they are completely identical.

carriers. All other virions will not leave "descendants", which significantly narrows the evolutionary corridor. The works of (Day et al., 2020) and (Markov et al., 2023) addressed a number of important issues regarding the epidemiology and evolution of the SARS-CoV-2 virus, including the mechanism of the emergence of recombinant strains.

Materials and methods

The most rational way to obtain both fast data processing speed and unlimited capabilities (which can be expanded if necessary) for their analysis, in our opinion, is to work with source FASTA files using the software package that combines our own software with third party libraries and programs. To date, a prototype that includes the minimum required functionality has been implemented. For the development, we used the C++ programming language available in Microsoft Visual Studio Community 2019. The hardware used was a PC based on an Intel Core i7-10700K processor, 3.8 GHz, 8 cores, 16 GB of RAM.

The methods used in this work mainly belong to the following two categories:

– theoretical estimates and numerical calculations of some important characteristics of the system under consideration, including the quality and reliability of the data;
– analysis of available genetic data using our own and existing software tools.

Whole genome genetic sequences of SARS-CoV-2. The GISAID and the Genbank databases provide, through a web interface, some functionality for studying the properties of the sequences they contain, but they are not flexible enough to perform the analysis required for investigation of the space of variants of SARS-CoV-2 genetic sequences, which is the goal of this work. There is also an API (Application Programming Interface) for GISAID (Wirth, Duchene, 2022), implemented in the R language. However, its capabilities also have limitations (including speed of operation with significant volumes of processed data) compared to direct access to genetic sequences stored as FASTA-files on a local workstation. GISAID significantly limits the possibilities of downloading from its website: no more than 2000 sequences per download, which completely eliminates the possibility of downloading a significant amount of data

"manually". The NCBI SARS-CoV-2 Data Hub has no such restrictions.

To analyze the already realized genetic variants of SARS-CoV-2, full-genome sequences from the GISAID (https://gisaid.org/) (Khare at al., 2021) and NCBI Virus SARS-CoV-2 Data Hub (https://www.ncbi.nlm.nih.gov/labs/virus/) (Sayers et al., 2022) were used. Sequences from Genbank (2019–2020) were downloaded to a local workstation and analyzed using our own software developed for this purpose, named ParSeq. Because of the limitations, sequences from GISAID were not downloaded – instead we accessed them through API to obtain only some of their properties (for example, full lengths of sequences; however, we were unable to obtain viral RNAs translatable part length and the positions of its start and end).

To calculate the edit distance between pairs of SARS-CoV-2 sequence variants (including a separate calculation of the number of mutations, deletions and insertions), the Nextstrain web resource (https://clades.nextstrain.org) was used.

Results

The estimation of the number of realized and potentially possible genetic variants of SARS-CoV-2 sequences

Let's start with considering the space of genetic sequences from a mathematical point of view, in the most general case. Any pair of sequences can be characterized by a measure of the difference between them, called the Levenshtein distance, or edit distance – the minimum number of point (single) substitutions (mutations, deletions, insertions) that must be made in the first sequence in order to transform it into the second one. Each element of the set of sequences of a given length L has a distance between itself and the empty sequence (Ø) which is exactly equal to L. The number of variants of nucleotide sequences of length L equals 4^L. The number of possible single mutations in a sequence of length L equals $3 \cdot L$ (the nucleotide at each position can be replaced by any of the other three). Also, $3 \cdot L$ different single deletions and $3 \cdot (L+1)$ different single insertions are possible. All possible single deletions for all possible sequences of length L compose the set of all possible sequences of length $(L-1)$, with the number of variants equal to $4^{(L-1)}$. And all possible single insertions for all possible sequences of length L result in a set of all possible sequences of length $(L+1)$, with the number of variants equal to $4^{(L+1)}$.

Let's consider all possible variants of nucleotide sequences of length $L = 2$ (Fig. 1). The set of sequences of $L = 2$ is quite small, but even in this simple case a hypercube in 4D space (tesseract, with 16 vertices) is required to represent all of this set's elements. For a more complex case, $L = 4$, in a similar way, a 6-dimensional hypercube (hexeract) with 64 vertices can be used (however, its visualization, together with the signatures of nodes and edges, will be oversaturated with details and difficult to perceive). Nevertheless, it can be displayed, in some degree, on a 2D plane using one of the Gray codes variants (Mütze, 2023) (this theory is closely connected with

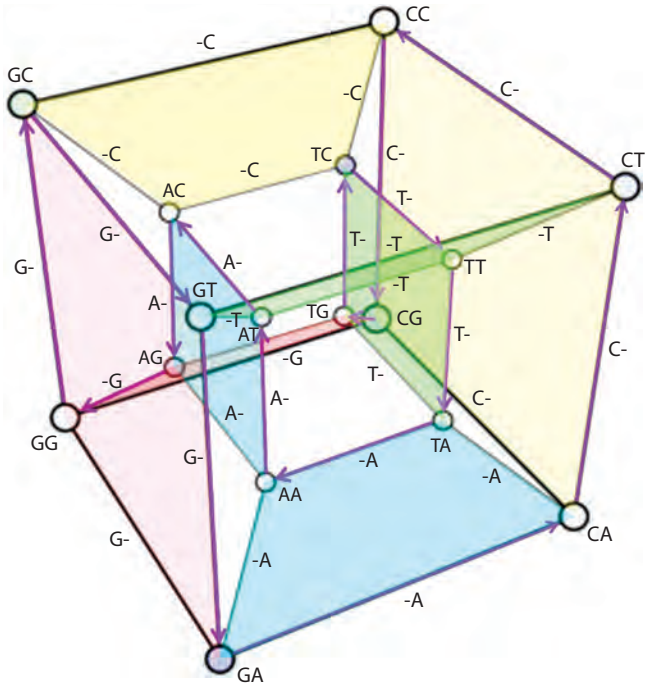


Fig. 1. The space of nucleotide sequence variants of length = 2, represented as a hypercube.

One of many Hamiltonian cycles on a hypercube (purple arrows) is presented – a closed path passing through each vertex exactly once. Each transition corresponds to a single change (mutation, deletion or insertion). There are also hyperplanes that can be associated with subsequences appearing from sequence of $L = 2$ after single deletions from the left (-A,-T,-G,-C) or from the right (A-,T-,G-,C-), which turn out to be the same in this simple case.

hypercubes), in this case – 2D code which we were able to find for this demonstration (Fig. 2).

The usual metric, such as the sum of squared differences of Cartesian coordinates, is apparently not suitable in this case.

The number of all possible sequences of equal length, in this case – the length of the reference genome of SARS-CoV-2, $L = 29903$, is very huge: 4^{29903}, or approximately $2.511 \cdot 10^{18003}$. In this space of variants, the set of sequences corresponding to the realized variants of the SARS-CoV-2 genome constitutes only a small part, composed of the point corresponding to the reference sequence and its small neighborhood, currently limited by the distance from the reference sequence to the most recent "Omicron" strain. It is possible to estimate the number of possible sequence variants within this distance. For the reference sequence, with $L = 29903$, the number of its variations with only one single mutation $= 3 \cdot L$, with two mutations $= (3 \cdot L)^2 - 3 \cdot L = 3 \cdot L \cdot (3 \cdot L - 1)$ (from all possible cases we subtract those in which second mutation occurs in the same position as the first one, and we get one of the already existing sequences – the reference one or a sequence which differs from it in only one position). Similarly, for the third mutation: $(3 \cdot L)^3 - ((3 \cdot L)^2 - 3 \cdot L)$, and so on. For $L = 29903$, the number of all variants of sequences with a number of

A = 00 T = 01 G = 10 C = 11					AA	GA	CA	TA	TT	AT	GT	CT	CC	TC	AC	GC	GG	CG	TG	AG
1	0	0	0	0	AA	AAAA	GAAA	CAAA	TAAA	TTAA	ATAA	GTAA	CTAA	CCAA	TCAA	ACAA	GCAA	GGAA	CGAA	TGAA	AGAA
2	1	0	0	0	GA	AAGA	GAGA	CAGA	TAGA	TTGA	ATGA	GTGA	CTGA	CCGA	TCGA	ACGA	GCGA	GGGA	CGGA	TGGA	AGGA
3	1	1	0	0	CA	AACA	GACA	CACA	TACA	TTCA	ATCA	GTCA	CTCA	CCCA	TCCA	ACCA	GCCA	GGCA	CGCA	TGCA	AGCA
4	0	1	0	0	TA	AATA	GATA	CATA	TATA	TTTA	ATTA	GTTA	CTTA	CCTA	TCTA	ACTA	GCTA	GGTA	CGTA	TGTA	AGTA
5	0	1	0	1	TT	AATT	GATT	CATT	TATT	TTTT	ATTT	GTTT	CTTT	CCTT	TCTT	ACTT	GCTT	GGTT	CGTT	TGTT	AGTT
6	0	0	0	1	AT	AAAT	GAAT	CAAT	TAAT	TTAT	ATAT	GTAT	CTAT	CCAT	TCAT	ACAT	GCAT	GGAT	CGAT	TGAT	AGAT
7	1	0	0	1	GT	AAGT	GAGT	CAGT	TAGT	TTGT	ATGT	GTGT	CTGT	CCGT	TCGT	ACGT	GCGT	GGGT	CGGT	TGGT	AGGT
8	1	1	0	1	CT	AACT	GACT	CACT	TACT	TTCT	ATCT	GTCT	CTCT	CCCT	TCCT	ACCT	GCCT	GGCT	CGCT	TGCT	AGCT
9	1	1	1	1	CC	AACC	GACC	CACC	TACC	TTCC	ATCC	GTCC	CTCC	CCCC	TCCC	ACCC	GCCC	GGCC	CGCC	TGCC	AGCC
10	0	1	1	1	TC	AATC	GATC	CATC	TATC	TTTC	ATTC	GTTC	CTTC	CCTC	TCTC	ACTC	GCTC	GGTC	CGTC	TGTC	AGTC
11	0	0	1	1	AC	AAAC	GAAC	CAAC	TAAC	TTAC	ATAC	GTAC	CTAC	CCAC	TCAC	ACAC	GCAC	GGAC	CGAC	TGAC	AGAC
12	1	0	1	1	GC	AAGC	GAGC	CAGC	TAGC	TTGC	ATGC	GTGC	CTGC	CCGC	TCGC	ACGC	GCGC	GGGC	CGGC	TGGC	AGGC
13	1	0	1	0	GG	AAGG	GAGG	CAGG	TAGG	TTGG	ATGG	GTGG	CTGG	CCGG	TCGG	ACGG	GCGG	GGGG	CGGG	TGGG	AGGG
14	1	1	1	0	CG	AACG	GACG	CACG	TACG	TTCG	ATCG	GTCG	CTCG	CCCG	TCCG	ACCG	GCCG	GGCG	CGCG	TGCG	AGCG
15	0	1	1	0	TG	AATG	GATG	CATG	TATG	TTTG	ATTG	GTTG	CTTG	CCTG	TCTG	ACTG	GCTG	GGTG	CGTG	TGTG	AGTG
16	0	0	1	0	AG	AAAG	GAAG	CAAG	TAAG	TTAG	ATAG	GTAG	CTAG	CCAG	TCAG	ACAG	GCAG	GGAG	CGAG	TGAG	AGAG

Fig. 2. The set of nucleotide sequence variants of length 4, depicted on a plane using 2D Gray codes.

The top edge of the table is coupled with the bottom, the left – with the right, i. e. one can map this set onto the surface of a torus. Then, when moving both horizontally and vertically (in the coordinate system of the table), in accordance with the properties of Gray codes, each pair of adjacent sequences will differ by exactly one (single) replacement (mutation).

mutations from 0 to n (relative to the reference sequence) is equal to $1.387 \cdot 10^{510}$ for $n = 103$, and for $L = 29847$ (56 deletions) – $1.108 \cdot 10^{510}$. Summing over all lengths from 29903 to 29847, we obtain $7.190 \cdot 10^{511}$.

Sequences with synonymous single nucleotide mutations that do not result in an amino acid change are also part of the total sequence variants space. However, the actual number of variants in the context of considering the structure and functions of proteins translated from viral RNA is significantly smaller due to the degeneracy of the genetic code (20 amino acids are encoded by 61 RNA triplets, i. e., on average, 3.05 triplets encode the same amino acid). Let's also take into account that not the entire genome of SARS-CoV-2 encodes proteins: 771 out of 29903 nucleotides are non-coding. As a result, the dependence proportional to $(3L)^n$ is transformed into $\approx((L–771)+(3 \cdot 771))^n$ and thus the corrected number of protein sequence variants can be estimated as $1.02 \cdot 10^{465}$. If we assume that someday the number of mutations will exceed the above-mentioned 103 pcs. by 10–11 times, then the sequence will most likely still be a coronavirus, but will already belong to a different species. For example, the bat coronavirus RaTG13, the closest neighbor of SARS-CoV-2 in the space of genetic sequence variants, differs from it by 1135 point mutations.

Let's try to look at the many variants of SARS-CoV-2 genetic sequences "tested" by nature from a biological point of view. The virus gets into a body (usually by airborne droplets, ending up in the lungs) and enters a cell, where a host ribosome begins to synthesize viral proteins in accordance with the nucleotide sequence of the SARS-CoV-2 genome. Among these proteins, there is a viral RNA polymerase (RdRp), which initiates a process of viral RNA replication. At the beginning, when there is only one viral RNA and one RdRp in the cell, the probability of their meeting is extremely low, but then, as these and other molecules accumulate in the cell, it starts to grow rapidly. As a result, the concentration reaches a level sufficient for the assembly of new virions, and when their number in the cell reaches approximately 10^5 pieces, these virions leave it and begin to infect neighboring cells, and more distant cells as well, if some of the virions enter the bloodstream and are distributed throughout the body. Considering that the number of viral particles in a patient's organism during the peak of infection can reach up to 10^{11} pcs. (Sender et al., 2021), let's divide this value by the average number of virions in an infected cell and get the number of infected cells in the body, 10^6. A human being has approximately $3 \cdot 10^{13}$ cells, so it appears that the percentage of infected among all is less than 10^{-4} %.

The frequency of errors occurrence during SARS-CoV-2 genome replication, according to (Amicone et al., 2022), is $1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ changes per position, per cell infection cycle, and is $(1–2) \cdot 10^{-6}$ according to (Markov et al., 2023), that is, approximately $1.4 \cdot 10^{-6}$ on average. Taking into account the length of the sequence, we obtain the probability of a single mutation occurring in the entire sequence per replication cycle ≈ 0.04. Thus, even if all infected cells in the body contain the same viral RNA variant at some moment, then after one replication cycle the body may contain all possible variants of single substitutions ($3 \cdot 29903$ pcs.) related to source viral RNA (which existed before the start of the cycle). So, there will be about 4 % of these (and most of them will not be viable), and 96 % will be exact copies of the replicated sequence. What will be the probability of occurrence of a viable non-synonymous mutation (changing not

only the RNA sequence of the virus, but also the amino acid sequence of one of its proteins), which is also superior to its predecessor in fitness? This question remains open; however, the required probability will definitely be very small. In the vast majority of cases, all copies of the virus spread by the infected person into the external environment are identical, and only rarely two variants occur simultaneously in one organism. How then new mutant variants not only appear, but also quickly displace their predecessors on a planetary scale every now and again?

Considering that the ratio of 4 % : 96 % with each subsequent replication cycle will change towards a decrease in the proportion of mutant sequences ("founder effect" (Ruan et al., 2020)) until they completely dissappear, we can suppose the following possible scenarios (with low probabilities) for the emergence and spread of mutant variants of SARS-CoV-2:

(a) The body does not have immunity to SARS-CoV-2 since it has not yet encountered it. A single copy of the viral RNA enters the cell; during the first round of replication, a mutation arises in it, and it turns out to be viable (this indeed can happen – with a low, but non-zero probability). Then all new virions synthesized by this cell will be carriers of this mutation, and if it is noticeably advantageous, they may have a chance of displacing the initial variant.

(b) The body already has immunity against SARS-CoV-2. It simultaneously contains two variants of SARS-CoV-2 virions – the one which is dominant in the population and the new one, mutant (arising by the mechanism from (a) or a recombinant). The immune system destroys the "old" variant that is familiar to it, but the new one goes unnoticed, passes through replication cycles and is transmitted further.

The probabilities of the occurrence of these two options have yet to be estimated, but even without this it is clear that the corridor of possible variants along which evolution took place turned out to be quite narrow. The opposite of this picture is, for example, the influenza virus, the distinctive feature and basis of survival of which is high variability due to the mechanisms of antigenic drift and antigenic shift (Kim et al., 2018).

We evaluate the modeling of the evolution of SARS-CoV-2 as possible, because despite the large number of variants that should have already been realized and which could have been realized from the point of view of mathematics (probability theory) and biology, in reality only a small part of them was realized and one can observe only a small part of the possible space of variants.

The development of the ParSeq software

To analyze the genetic sequences of SARS-CoV-2, we developed the software called ParSeq (**Par**ser of **Seq**uences) – parser and analyzer of SARS-CoV-2 nucleotide sequences in FASTA format, which we already used while working on analysis of the SARS-CoV-2 epidemic in regions of Siberia (Palyanova et al., 2023). Its main abilities already implemented at the moment are described below:

- Loading and parsing one or many FASTA files (using the list of file names) for further analysis, including the following data fields: full-genome nucleotide sequence, "Accession ID", "Length", "Pango lineage", "Nuc. completeness", "Collection date", "Geo location" and "Country".
- Primary analysis of the nucleotide sequence, including calculation of its length and nucleotide content (A, U(T), G, C and non-identified nucleotides represented by the letter "N"). Also, in some sequences, the following letters of the extended alphabet are found sometimes: (https://www.bioinformatics.org/sms/iupac.html):

R	Y	S	W	K
A ‖ G	C ‖ T	G ‖ C	A ‖ T	G ‖ T
M	B	D	H	Y
A ‖ C	C ‖ G ‖ T	A ‖ G ‖ T	A ‖ C ‖ T	A ‖ C ‖ G

- Determination of the positions of the beginning and end of the coding part of the sequence. In the case of a reference sequence, its total length is 29903 nt, the length of non-coding 5′ UTR – 265 nt, non-coding 3′ UTR – 229 nt. To do this, the following simple algorithm is used: in the case of a 5′ UTR, we move along the sequence from its beginning to the 500th nucleotide (for convenience, a "round" value was chosen, for which 265 is approximately in the middle) with a window of length 17 and count the number of nucleotide matches in this window with a fragment of the reference sequence of the same length, corresponding to the interval 266–282 (where 266 is the position of the translation start in the reference genome). If 14 or more out of 17 positions match, then the position is determined correctly (numerical parameters are defined as sufficient for correct operation in the vast majority of cases using a small window length – to avoid unnecessary calculations). In the case of a 3′ UTR, everything is similar – with a 17 nt long window we move along the last 500 nucleotides of the analyzed sequence, comparing its contents with the 17 nucleotides that end the coding region of the reference sequence. The criterion of the correct position is the same – 14 or more matches within the 17 nt long window.
- Calculation of the lengths of the non-coding 5′ UTR and 3′ UTR, as well as the coding region located between them, which makes up the vast majority of the genome of the viral sequence (98.35 % of its length in the case of the reference sequence).
- Calculation of distributions of these values for any selection of SARS-CoV-2 genome sequences (e. g., within a specified time interval for the collection date, or for sequences containing no more than a specified number of "N"s, etc.; combinations of various filters are also supported).
- Calculation of distribution of sequences by number of their lengths.

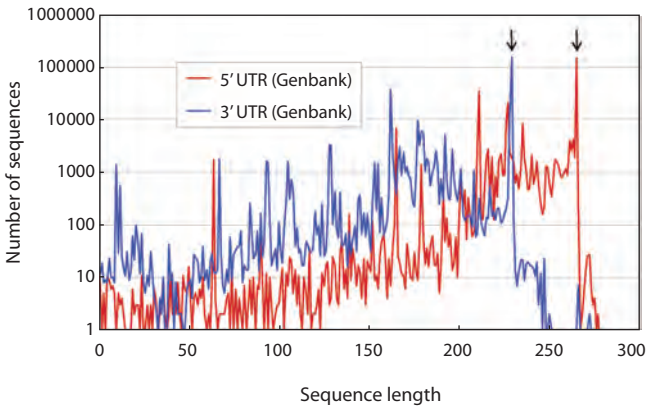


Fig. 3. Distributions of 5′ UTR and 3′ UTR lengths for sequences from the Genbank database for the period from the emergence of SARS-CoV-2 (at the end of 2019) to the end of 2020.

The lengths of 5′ UTR and 3′ UTR in the reference genome of SARS-CoV-2 are 265 and 229 nt, respectively. The peak values of both curves correspond precisely to these lengths.

The results obtained using ParSeq software

Using the software we developed, we analyzed the nucleotide sequences of SARS-CoV-2, available to users around the world thanks to the Genbank and GISAID projects. As a result, the following facts were established.

1. The calculation of the distribution of genetic sequences by their full lengths (5′ UTR + coding sequence + 3′ UTR) among sequences with a length ≥ 28000 revealed that for data from Genbank (for the period from 01.12.2019 to 31.12.2022) the minimum length of the complete sequence was 28784, and the maximum was 29985. The vast majority of the distribution corresponds to lengths less than or equal to the reference sequence length, 29903. The difference between the reference and the minimum length was 1119. This does not match well with the data from the Table, according to which the maximum difference between the length reference and any other sequence is about 159 (103 mutations + 56 deletions). Moreover, with such a difference, this sequence would most likely belong to a different type of virus, since the reference sequence of SARS-CoV-2 and the bat coronavirus RaTG13 have a similar difference (GenBank MN996532.2, collection_date=24-Jul-2013). According to (Li et al., 2023), they differ by 96.2 %, i. e. by 1136 single mutations (distributed throughout the sequence). Calculation of the distance between the same sequences, made using the Nextstrain web service, showed a difference of 1135 single mutations, as well as 20 deletions (in the coding part of RaTG13 relative to the reference sequence of SARS-CoV-2). The total genome length of RaTG13 is 28855, i. e. the number of deletions relative to SARS-CoV-2 is 48. Most probably, such too short or too long sequences correspond to low-quality data with errors in genome assembly.

Because the difference between the full length of the SARS-CoV-2 reference genome and the rest of the sequences stored in the database for some of them significantly exceeds the number of differences (point mutations, deletions and insertions) between the SARS-CoV-2 reference genome and the most different variant of "Omicron" (see the last row in the Table), we decided to study the distribution not only of

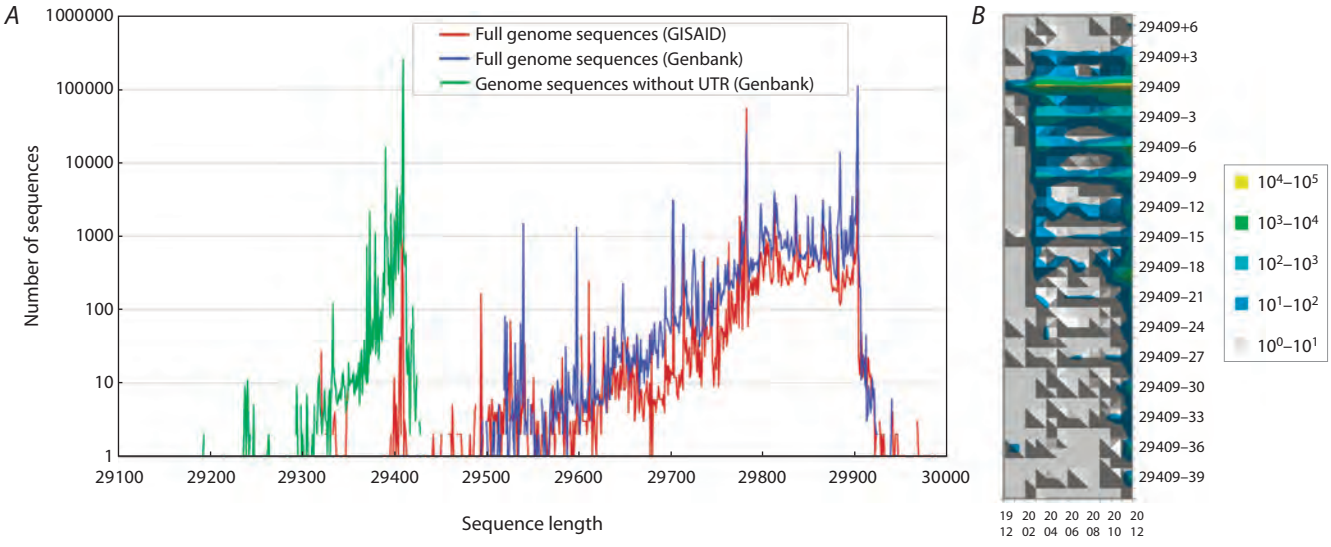


Fig. 4. *A*, Distribution of lengths of full genomes (GISAID, Genbank) and their lengths without UTRs (Genbank) during a period from the emergence of SARS-CoV-2 in 2019 until the end of 2020. The full length of the SARS-CoV-2 reference genome is 29903 nt, and the length of its coding part (without UTRs) is 29409. Peak (maximal) values for all three curves correspond to these values. *B*, Change of the lengths of the SARS-CoV-2 genomes coding part (Genbank) during 12.2019–12.2020 by months. Horizontal signatures are numerical representations of the year and the month, vertical represent the lengths of the genome coding part; colors correspond to the frequency of genome sequences with a specified length (logarithmic scale).

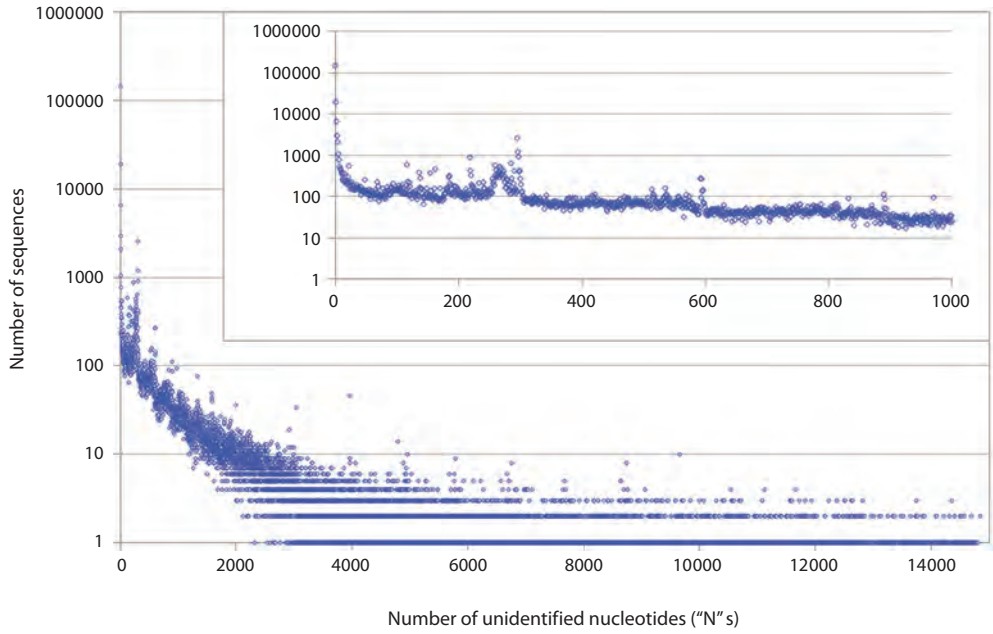


Fig. 5. Distribution of SARS-CoV-2 sequences by the number of non-identified or partially identified nucleotides in the translatable part of their genomes (from Genbank, collection date within the interval from 12.2019 until 12.2020).

The inset contains part of the same graph as in the main picture, but for the area from 0 to 1000 horizontally.

full lengths of genomes, but of their coding and non-coding regions as well (Fig. 3, 4). As seen in Figure 3, the 5′ UTR and 3′ UTR regions found in the databases have lengths from 0 to reference values, and in a small number of cases they are slightly longer. Sequences, the 5′ UTR and 3′ UTR lengths of which coincide with the reference ones, account for 49.7 and 51.2 % of their total number, respectively. Sequences, the 5′ UTR and 3′ UTR lengths of which differ from the reference ones by no more than 10 nt, constitute 55.9 and 55.7 % of their total number, respectively.

Also, Figure 4, *A* shows that the main source of the observed scatter in the distribution of full lengths of the SARS-CoV-2 genomes was indeed due to the scatter in lengths of the untranslated regions – 5′ and 3′ UTRs. If we consider only the coding part, the scatter is significantly reduced: 84.9 % of all sequences have the length of the coding part equal to the length of the reference genome, and 90.7 % have a length of the coding part that differs from it by no more than 10 nt. In addition, Figure 4, *B* shows that among the genomes, the length of the coding part of which differs from that of the reference sequence (29409), prevail those in which this difference is a multiple of 3 – to prevent a shift in the reading frame during translation, which usually leads to non-viability. Thus, most of the processed viral sequences appear to be biologically meaningful.

It can be seen that the distributions obtained based on complete genomes data from GISAID (obtained using the access through API) and Genbank (through analysis of downloaded sequences using ParSeq software) have a fairly high similarity – probably due to the fact that most sequences are stored in both databases (see Fig. 4). The question about how many sequences that differ in length from the reference one actually have deletions or insertions, and how many of them have these differences due to errors in sequencing and genome assembly, remains open.

2. When studying genetic sequences representing the genomes of different variants of a virus that change over time, there is often a need to compare them with each other. Even if a pair of sequences have identical coding region lengths, the ability to calculate the amount of difference between them (the number of point mutations) will depend on whether the sequences contain undefined nucleotides, usually denoted "N", or letters other than the standard A, T(U), G and C. Using the ParSeq software and the genomes of SARS-CoV-2 sequences collected in 2019–2020 (from the Genbank database), we calculated the distribution of sequences by the number of unidentified or partially identified nucleotides in them (Fig. 5).

Throughout most of the graph, the number of sequences decreases exponentially with the number of unidentified nucleotides, although there are areas with some peculiarities. The number of sequences for which all nucleotides are identified is 47.8 %, the number of sequences where less than 10 nucleotides are uncertain is 58.9 %. Thus, for the analysis of evolutionary changes occurring in the SARS-CoV-2 virus, a significant part of the total number of sequences is suitable.

Discussion

We carried out a number of estimates, calculations and computational analyses (using software developed by us), to improve our understanding of the space of SARS-CoV-2 genetic sequences variants, find out what are its main properties and features associated with a quite long genomic sequence (for RNA viruses) and a low frequency of mutations occurring in the process of its replication.

There are viruses the genome of which is significantly smaller than that of SARS-CoV-2. Because of its relatively large length, the number of viable variants exceeds that of small viruses. Let's try to determine some other landmarks in the space of viral genetic sequence variants. SARS-CoV-2 belongs to single-stranded RNA(+) viruses (Modrow et al., 2013). One on the smallest ssRNA(+) human viruses is the Astrovirus type 1 (genome length = 6771 nt) (Lewis et al., 1994). An even smaller ssRNA(+) genome (4294 nt) belongs to the shrimp nodavirus (*Penaeus vannamei nodavirus*) (Chen et al., 2019). The total number of variants of different sequences of these two lengths is equal to $3.533 \cdot 10^{4076}$ and $1.760 \cdot 10^{2585}$, correspondingly.

If in our search for the smallest viral genome we consider DNA viruses as well, then among the record holders we will find pig circovirus type 1, *Porcine circovirus* 1 (PCV-1) (Cao et al., 2018), with genome size equal to 1757–1759 bp (17 times less than that of SARS-CoV-2). The number of possible variants of genetic sequences of such length is $6.597 \cdot 10^{1057}$. This is still a far cry from the number of variants that were potentially available to SARS-CoV-2 during the period of its existence (3.5 years), $7.985 \cdot 10^{511}$. And a genome with a length of 850 nt would have a very close number of possible sequence variants, $5.636 \cdot 10^{511}$. However, there are single-stranded circular RNA infectious agents with even shorter sequence lengths (from 246 to 467 nt), named viroids (Katsarou et al., 2015). Their RNA is not protected by any envelope and does not encode proteins.

So, SARS-CoV-2, like all other viruses, potentially has a very large number of possible variants, compared both to the number of collected and sequenced specimens, and to the number of variants that have been "tested" during evolution, but turned out to be non-viable.

And finally let's get back to the bat coronavirus RaTG13 (L = 29855) – the nearest neighbor of SARS-CoV-2 in the space of genetic sequences variants, which differ from it by 1135 single mutations. The total number of variants of sequences generated by the reference SARS-CoV-2 genome modified by a number of mutations (from 1 to 1135), may be estimated as $\approx 2.943 \cdot 10^{5621}$, which exceeds by many orders of magnitude the total number of possible variants of sequences as long as 4294 nt ($1.76 \cdot 10^{2585}$) and 6771 nt ($3.53 \cdot 10^{4076}$), i. e. it can contain in itself the amount of information enough for a huge number of different small viruses.

The global phylogenetic tree of the SARS-CoV-2 shows that the virus cannot remain unchanged over time; it is forced to alter, apparently due to the fact that natural selection pressure acts on it. Another reason for changes is intraspecific competition – for example, variants with faster RNA polymerases displace variants with slower ones (since their number grows faster) and thereby reduce the incubation period of the virus over time; less lethal strains allow the virus to spread longer and wider (the carrier remains alive and spreads the virus throughout almost the entire period of the disease; an infected person with mild symptoms or their absence remains socially active and infects more people in their environment). Unlike the viroids mentioned above, changes in the genome of real viruses, including SARS-CoV-2, can have different effects on intraspecific competition depending on the functions of the proteins encoded in the genome. This issue remained outside the scope of this work, but in subsequent publications we plan to pay due attention to it.

In addition, the formation of immunity to this virus in humanity also has an impact on further virus evolution, and there are probably other mechanisms too. Moreover, all these changes should occur without compromising the functionality of the virus. Thus, it turns out that the space of variants available to the SARS-CoV-2 coronavirus is quite narrow, and the trajectories of its development may be determined to some extent. Indeed, the SARS-CoV-2 genome has been shown to have a much lower mutation rate and genetic diversity compared to the SARS-CoV virus that caused the atypical pneumonia outbreak in 2002–2003 (Jia et al., 2020; Zhou et al., 2020; Nikonova et al., 2021). Thus, for example, for the SARS-CoV-2 S-protein, the d_N and d_S values appeared to be approximately 12 and 7 times lower than those for SARS-CoV (where d_N is the fraction of sequences in a sample of genomes that contain non-synonymous mutations in a particular gene; d_S is a similar value, but for synonymous mutations). For more conservative genes, ORF1a and ORF1b, the ratios of mutation frequencies

$$(d_N^{\text{SARS-CoV-2}}/d_N^{\text{SARS-CoV}},\ d_S^{\text{SARS-CoV-2}}/d_S^{\text{SARS-CoV}})$$

are less than those for S-protein, but values for SARS-CoV-2 are also lower than the corresponding values for SARS-CoV (belonging to the interval from $\frac{1}{4.8}$ to $\frac{1}{1.5}$). The hypothesis about the partial determinism of coronavirus evolutionary trajectories is that if the development of the SARS-CoV-2 pandemic, from its very beginning in December 2019, due to random factors, would have gone somewhat differently, then, despite this, sooner or later, in the same order or in a different one, the space of viable variants "visited" by the virus would still be approximately the same. The above allows to suggest that creating an evolutionary simulator based on an analysis of the trajectories of virus change over time might be quite possible, which is part of our future plans.

Conclusion

Investigation of the space of genetic sequence variants is an important step in developing approaches for modeling the evolution of viruses and other organisms. To build a new, significantly more realistic model of virus evolution, capable of calculating potentially possible viral genome sequences

variants, which are not yet realized in nature, in order to proactively prevent their emergence, it is necessary to answer questions such as: What is the probability of recombination and are there preferred positions in which it usually occurs? Can we guess or calculate which variant will be realized and which will not be viable? Could "Delta" or "Omicron" genetic sequences have been predicted (calculated before their emergence)? And finally, if it were possible to create a realistic model of the evolution of SARS-CoV-2 and calculate the process several times from the very beginning, from the initial reference sequence, would it proceed differently each time and lead to significantly different results, or would everything happen approximately the same with minor variations?

References

Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Software*. 2021;6(67):3773. DOI 10.21105/joss.03773

Amicone M., Borges V., Alves M.J., Isidro J., Zé-Zé L., Duarte S., Vieira L., Guiomar R., Gomes J.P., Gordo I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health*. 2022;10(1):142-155. DOI 10.1093/emph/eoac010

Campagnola G., Govindarajan V., Pelletier A., Canard B., Peersen O.B. The SARS-CoV nsp12 polymerase active site is tuned for large-genome replication. *J. Virol.* 2022;96(16):e0067122. DOI 10.1128/jvi.00671-22

Cao L., Sun W., Lu H., Tian M., Xie C., Zhao G., Han J., Wang W., Zheng M., Du R., Jin N., Qian A. Genetic variation analysis of PCV1 strains isolated from Guangxi Province of China in 2015. *BMC Vet. Res.* 2018;14(1):43. DOI 10.1186/s12917-018-1345-z

Chen N.C., Yoshimura M., Miyazaki N., Guan H.-H., Chuankhayan P., Lin C.-C., Chen S.-K., Lin P.-J., Huang Y.-C., Iwasaki K., Nakagawa A., Chan S.I., Chen C.J. The atomic structures of shrimp nodaviruses reveal new dimeric spike structures and particle polymorphism. *Commun. Biol.* 2019;2:72. DOI 10.1038/s42003-019-0311-z

Day T., Gandon S., Lion S., Otto S.P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* 2020;30(15):R849-R857. DOI 10.1016/j.cub.2020.06.031

Grebennikov D., Kholodareva E., Sazonov I., Karsonova A., Meyerhans A., Bocharov G. Intracellular life cycle kinetics of SARS-CoV-2 predicted using mathematical modelling. *Viruses*. 2021;13(9):1735. DOI 10.3390/v13091735

Jia Y., Shen G., Nguyen S., Zhang Y., Huang K., Ho H., Hor W., Yang C., Bruning J.B., Li C., Wang W. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv*. 2020. DOI 10.1101/2020.04.09.034942

Karimzadeh S., Raj B., Nguyen T.H. Review of infective dose, routes of transmission and outcome of COVID-19 caused by the SARS-COV-2: comparison with other respiratory viruses. *Epidemiol. Infect.* 2021;149:e96. DOI 10.1017/S0950268821000790

Katsarou K., Rao A.L.N., Tsagris M., Kalantidis K. Infectious long non-coding RNAs. *Biochimie*. 2015;117:37-47. DOI 10.1016/j.biochi.2015.05.005

Khare S., Gurry C., Freitas L., Schultz M.B., Bach G., Diallo A., Akite N., Ho J., Lee R.T., Yeo W., Curation Team GC, Maurer-Stroh S. GISAID's role in pandemic response. *China CDC Weekly*. 2021;3(49):1049-1051. DOI 10.46234/ccdcw2021.255

Kim H., Webster R.G., Webby R.J. Influenza virus: dealing with a drifting and shifting pathogen. *Viral Immunol.* 2018;31(2):174-183. DOI 10.1089/vim.2017.0141

Lewis T.L., Greenberg H.B., Herrmann J.E., Smith L.S., Matsui S.M. Analysis of astrovirus serotype 1 RNA, identification of the viral RNA-dependent RNA polymerase motif, and expression of a viral structural protein. *J. Virol.* 1994;68(1):77-83. DOI 10.1128/JVI.68.1.77-83.1994

Li P., Hu J., Liu Y., Ou X., Mu Z., Lu X., Zan F., Cao M., Tan L., Dong S., Zhou Y., Lu J., Jin Q., Wang J., Wu Z., Zhang Y., Qian Z. Effect of polymorphism in *Rhinolophus affinis* ACE2 on entry of SARS-CoV-2 related bat coronaviruses. *PLoS Pathog.* 2023;19(1):e1011116. DOI 10.1371/journal.ppat.1011116

Malone B., Urakova N., Snijder E.J., Campbell E.A. Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat. Rev. Mol. Cell Biol*. 2022;23(1):21-39. DOI 10.1038/s41580-021-00432-z

Markov P.V., Ghafari M., Beer M., Lythgoe K., Simmonds P., Stilianakis N.I., Katzourakis A. The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 2023;21(6):361-379. DOI 10.1038/s41579-023-00878-2

Modrow S., Falke D., Truyen U., Schätzl H. Viruses with single-stranded, positive-sense RNA genomes. In: Molecular Virology. Berlin: Springer, 2013;185-349. DOI 10.1007/978-3-642-20718-1_14

Mütze T. Combinatorial Gray codes – an updated survey. *Electron. J. Comb.* 2023;30(3):DS26. DOI 10.37236/11023

Nikonova A.A., Faizuloev E.B., Gracheva A.V., Isakov I.Yu., Zverev V.V. Genetic diversity and evolution of the biological features of the pandemic SARS-CoV-2. *Acta Naturae*. 2021;13(3):77-89. DOI 10.32607/actanaturae.11337

Palyanova N., Sobolev I., Alekseev A., Glushenko A., Kazachkova E., Markhaev A., Kononova Y., Gulyaeva M., Adamenko L., Kurskaya O., Bi Y., Xin Y., Sharshov K., Shestopalov A. Genomic and epidemiological features of COVID-19 in the Novosibirsk region during the beginning of the pandemic. *Viruses.* 2022;14(9):2036. DOI 10.3390/v14092036

Palyanova N.V., Sobolev I.A., Palyanov A.Y., Kurskaya O.G., Komissarov A.B., Danilenko D.M., Fadeev A.V., Shestopalov A.M. The development of the SARS-CoV-2 epidemic in different regions of Siberia in the 2020–2022 period. *Viruses*. 2023;15:2014. DOI 10.3390/v15102014

Ruan Y., Luo Z., Tang X., Li G., Wen H., He X., Lu X., Lu J., Wu C.I. On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? *Natl. Sci. Rev.* 2020;8(1):nwaa246. DOI 10.1093/nsr/nwaa246

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26. DOI 10.1093/nar/gkab1112

Sender R., Bar-On Y.M., Gleizer S., Bernshtein B., Flamholz A., Phillips R., Milo R. The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci. USA*. 2021;118(25):e2024815118. DOI 10.1073/pnas.2024815118

Shannon A., Selisko B., Le N.T., Huchting J., Touret F., Piorkowski G., Fattorini V., Ferron F., Decroly E., Meier C., Coutard B., Peersen O., Canard B. Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. *Nat. Commun*. 2020;11(1):4682. DOI 10.1038/s41467-020-18463-z

Sonnleitner S.T., Prelog M., Sonnleitner S., Hinterbichler E., Halbfurter H., Kopecky D.B.C., Almanzar G., Koblmüller S., Sturmbauer C., Feist L., Horres R., Posch W., Walder G. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun.* 2022;13(1):2560. DOI 10.1038/s41467-022-30163-4

Wirth W., Duchene S. GISAIDR: programmatically interact with the GISAID databases. *Zenodo*. 2022. DOI 10.5281/zenodo.6474693

Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y., Tao Z.W., Tian J.H., Pei Y.Y., Yuan M.L., Zhang Y.L., Dai F.H., Liu Y., Wang Q.M., Zheng J.J., Xu L., Holmes E.C., Zhang Y.Z. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. DOI 10.1038/s41586-020-2008-3

Xu H., Zhang Y., Yuan M., Ma L., Liu M., Gan H., Liu W., Lum G.G.A., Tao F. Basic reproduction number of the 2019 novel coronavirus disease in the major endemic areas of China: a latent profile analysis. *Front. Public Health*. 2021;9:575315. DOI 10.3389/fpubh.2021.575315

Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., Si H.R., Zhu Y., Li B., Huang C.L., Chen H.D., Chen J., Luo Y., Guo H., Jiang R.D., Liu M.Q., Chen Y., Shen X.R., Wang X., Zheng X.S., Zhao K., Chen Q.J., Deng F., Liu L.L., Yan B., Zhan F.X., Wang Y.Y., Xiao G.F., Shi Z.L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. DOI 10.1038/s41586-020-2012-7

ORCID ID

A.Yu. Palyanov orcid.org/0000-0003-1108-1486
N.V. Palyanova orcid.org/0000-0002-1783-5798

Funding. This research was funded by RSF, grant number 23-64-00005.

Acknowledgements. We gratefully acknowledge all data contributors, i. e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing them via the GISAID Initiative and Genbank SARS-CoV-2 Data Hub, on which this research is based. We are also grateful to the Authors of the Nextclade project, which provides online tools for analysis and visualization of genetic data on various viruses.

Conflict of interest. The authors declare no conflict of interest.

Received July 16, 2023. Revised September 14, 2023. Accepted September 18, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):851-858
DOI 10.18699/VJGB-23-98

Original Russian text https://vavilovj-icg.ru/

Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data

X. Fu[1], S.S. Tamozhnikov[2], A.E. Saprygin[2, 3], N.A. Istomina[1], D.I. Klemeshova[3], A.N. Savostyanov[1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[3] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
a-sav@mail.ru

Abstract. The development of objective methods for assessing stress levels is an important task of applied neuroscience. Analysis of EEG recorded as part of a behavioral self-control program can serve as the basis for the development of test methods that allow classifying people by stress level. It is well known that participation in meditation practices leads to the development of skills of voluntary self-control over the individual's mental state due to an increased concentration of attention to themselves. As a consequence of meditation practices, participants can reduce overall anxiety and stress levels. The aim of our study was to develop, train and test a convolutional neural network capable of classifying individuals into groups of practitioners and non-practitioners of meditation by analysis of event-related brain potentials recorded during stop-signal paradigm. Four non-deep convolutional network architectures were developed, trained and tested on samples of 100 people (51 meditators and 49 non-meditators). Subsequently, all structures were additionally tested on an independent sample of 25 people. It was found that a structure using a one-dimensional convolutional layer combining the layer and a two-layer fully connected network showed the best performance in simulation tests. However, this model was often subject to overfitting due to the limitation of the display size of the data set. The phenomenon of overfitting was mitigated by changing the structure and scale of the model, initialization network parameters, regularization, random deactivation (dropout) and hyperparameters of cross-validation screening. The resulting model showed 82 % accuracy in classifying people into subgroups. The use of such models can be expected to be effective in assessing stress levels and inclination to anxiety and depression disorders in other groups of subjects.
Key words: convolutional neural networks; EEG; event-related brain potentials; meditation; stop-signal paradigm.

For citation: Fu X., Tamozhnikov S.S., Saprygin A.E., Istomina N.A., Klemeshova D.I., Savostyanov A.N. Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):851-858. DOI 10.18699/VJGB-23-98

Сверточные нейронные сети для классификации по данным ЭЭГ здоровых людей, практикующих или не практикующих медитацию

С. Фу[1], С.С. Таможников[2], А.Е. Сапрыгин[2, 3], Н.А. Истомина[1], Д.И. Клемешова[3], А.Н. Савостьянов[1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[3] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
a-sav@mail.ru

Аннотация. В настоящее время разработка объективных методик для оценки уровня стресса является чрезвычайно актуальной задачей прикладной нейронауки. Анализ электроэнцефалограммы (ЭЭГ), записанной в условиях выполнения заданий на самоконтроль поведения, может служить основой для разработки тестовых методик, позволяющих классифицировать людей по уровню стресса. Хорошо известно, что одним из следствий медитационной практики является выработка у участников навыков произвольного контроля над собственным ментальным состоянием за счет повышенной концентрации внимания на самом себе. На фоне медитационной практики часто происходит снижение общего уровня тревожности и стресса. Целью нашего исследования было разработать, обучить и протестировать сверточную нейронную сеть, способную классифицировать людей на группы участвующих или не участвующих в медитационной практике на основе анализа вызванных потенциалов головного мозга, записанных при выполнении заданий парадигмы стоп-сигнал. Были разработаны четыре архитектуры неглубоких сверточных сетей, которые были обучены и протестированы на выборке из 100 человек (51 медитатор и 49 не-медитатор). В дальнейшем все структуры были дополнительно протестированы на независимой выборке в 25 человек. Установлено, что структура, использующая одномерный сверточный слой, который объединяет слой и двуслойную полностью подключенную сеть, показала наилучшие результаты

работы в имитационных тестах. Однако эта модель была часто подвержена переобучению из-за ограничения размера отображения набора данных. Явление переобучения было смягчено при помощи изменения структуры и масштаба модели, параметров сети инициализации, регуляризации, случайной деактивации (dropout) и гиперпараметров скрининга перекрестной проверки. В итоге нами получена модель, которая показала 82 % точность в классификации людей на подгруппы. Можно ожидать, что использование таких моделей окажется эффективным методом для оценки уровня стресса и предрасположенности к тревожным и депрессивным расстройствам в других группах испытуемых.
Ключевые слова: сверточные нейронные сети; ЭЭГ; вызванные потенциалы мозга; медитация; парадигма стоп-сигнал.

Introduction

Stress is one of the most common problems in modern society, and the search for effective methods to assess stress levels is important for early detection of the risk of mental and psychosomatic disorders (Kuh et al., 2003; Kuznetsova et al., 2016). Most psychological methods of assessing stress levels are based on the use of questionnaires, in which the respondent answers questions about their subjective mental condition. The weak point of this approach is the high probability of incorrect self-assessments, arising either from a person's unwillingness to report their problems, or as a result of a low ability to recognize changes in their own condition (Iwata, Higuchi, 2000; McCrae et al., 2000). A possible solution to this problem is to develop objective approaches to the diagnosis of mental traits or conditions based on the analysis of brain signals, such as fMRI or EEG.

Meditation is a system of special mental practices aimed at establishing voluntary self-control over one's mental state. Although meditation initially appears as part of religious practices, especially common in oriental religions, at present this phenomenon is a popular topic of interest among scientific researchers. Meditation is considered as a basis for the creation of non-invasive, non-drug techniques that reduce the risk of a wide range of mental or psychosomatic diseases. A number of studies have shown that meditation has many positive effects on mental health, including a general reduction in stress and the level of propensity to depression (Chiesa et al., 2011; Saeed et al., 2019). The analysis of the EEG recorded during recognition of emotional stimuli revealed significant effects of meditation on the state of the human brain (Aftanas, Golosheykin, 2005; Atchley et al., 2016; Savostyanov et al., 2020). Therefore, the comparison of EEG in practitioners and non-practitioners of meditation can be considered as an experimental model that allows the development of methods for assessing stress levels.

Stop-signal paradigm (SSP) is an experimental method for evaluating an individual's ability for voluntary self-control of their own movements in a changing external environment (Logan, Cowan, 1984; Band et al., 2003). The SSP allows us to assess the balance of two processes – activation and inhibition of behavior under conditions of insufficient time for making a decision. A number of studies have shown that SSP is an effective method for diagnosing the level of personal anxiety and propensity to depression (Hsieh et al., 2021; Zelenskih et al., 2022). It can be assumed that the dynamics of brain activity during SSP will serve as a marker distinguishing practitioners and non-practitioners of meditation from each other.

Artificial neural network is a developing technology based on machine learning, which is widely used in various fields. Compared to other traditional methods of machine classification, such as linear discriminant analysis and the k-nearest neighbor algorithm, artificial neural networks provide more accurate results of classifying individuals according to their behavioral and neurophysiological characteristics (Khosla et al., 2020). Therefore, in comparison with the support vector machine, an artificial neural network is better suited for the tasks of multiple classification, providing convenience for further research, as well as more efficient fitting of complex nonlinear relationships.

The purpose of our research is to develop, train and test an artificial neural network that allows, based on the analysis of event-related brain potentials in the stop-signal paradigm, to classify individuals according to the criterion of whether they practice meditation. We assume that afterwards the neural network created in this way will be able to assess individual level of stress and propensity to anxiety-depressive disorders.

Methods of experimental research

Participants. A group of people practicing samadhi meditation (also called "mindfulness meditation") was examined in July–August 2018 on the premises of the Baikal Retreat Center (http://www.geshe.ru/). The experimental group included 51 healthy, right-handed participants from 25 to 66 years old (32 men; average age = 41.0, SD = 8.3), practicing meditation for a period of 5 to 15 years. The control group was examined in October–November 2019 on the premises of the medical college of the village of Khandyga, Tomponsky district of the Republic of Sakha (Yakutia). The control group included 49 healthy, right-handed participants from 22 to 58 years old (22 men; average age = 38.0, SD = 8.3) who had never participated in meditation or yoga practices.

The protocol of the study was approved by the local Ethics Committee of the Research Institute of Neurosciences and Medicine in accordance with the Helsinki Declaration of Biomedical Examinations. All the participants signed informed consent to participate in the surveys.

Experimental procedure. The experiment was organized on the basis of the stop-signal paradigm proposed in 1984 (Logan, Cowan, 1984) and modified by A.N. Savostyanov and co-authors (Savostyanov et al., 2009). The experiment was organized in the form of the computer interactive game "Hunt". One of two images appeared on the computer screen:

a deer, or a tank. The participant had to press the keyboard button corresponding to the picture. The response time was limited to 0.75 seconds. If the participant pressed the button correctly and faster than 0.75 seconds, their game score increased. If the participant pressed the buttons incorrectly or reached a time out, then their game score decreased.

In total, 135 stimuli were presented to each participant. In 35 cases, after the onset of the target signal, a stop-signal was presented (a red square with the inscription "Stop"), which meant that the participant had to interrupt the movement that had already begun. If the participant did not press the button after the stop-signal, their score did not change. If the participant pressed the button after the stop-signal, their score decreased. The order of activation and stopping trials was randomized. The sequence of "deer" and "tank" stimuli was also randomized. The interval between the end of the previous task and the start of a new task varied from 3 to 7 seconds. The total duration of the experiment was approximately 12 minutes.

Preprocessing of experimental data. EEG rejection of artifacts was done by the ICA method (Delorme, Makeig, 2004). The initial EEG signal was filtered at 1–40 Hz and referenced to average of all channels. The data was epoched relative to the onset trigger of the target stimulus (deer or tank) at a time interval from –1 to +3 seconds. The baseline EEG level was set in the range from –1000 to –250 ms. In total, 80 to 90 EEG epochs were obtained for each participant, after exclusion of all the trials containing the stop-signal or artifacts. After excluding artifacts, event-related potentials (ERPs) were calculated separately for each EEG channel, averaged over all trials and all participants.

The ERP calculation was conducted in the ERPLAB toolbox for MATLAB. Amplitude-time ERP graphs were made for each EEG channel. Then a visual preview of the ERP graph for the C3 channel was performed. In this lead, the ERP motor peaks stand out the most. In particular, two peaks were selected for this lead – an early premotor peak, the amplitude of which precedes pressing the button (the so-called readiness potential) and a late motor peak, the amplitude of which reaches a maximum when the button is pressed. From viewing this visual, the time limits of both the early and late peak were established. After that, the amplitude in each of these time windows was calculated separately for each person and each EEG channel, but averaged over all trials of the activation condition of the task for each participant. The calculation of the averaged amplitude was made using the ERPLAB toolbox (https://erpinfo.org/erplab). The amplitude values were surveyed to the baseline level for each participant separately. The obtained values were used as training and test data for artificial neural networks.

EEG data acquisition. The general structure of the input data is shown in Figure 1. For each participant, EEG was analyzed for 64 channels located at different points of the head surface. According to the international scheme of 10–20 %, the name of the electrode reflects its spatial position. The initial EEG signal for each channel is presented as a continuous series of measurements of the potential difference between the surface electrode and the referent with a time resolution of 1,000 measurements per second.

ERP extraction. When calculating the ERP (event-related potential) amplitude, the researcher selects several time windows, in each of which all amplitude values are summed over all time points and averaged over all tests. The amplitude values in different windows reflect the temporal dynamics of the neurophysiological process. We selected two time windows (250–350 and 550–900 ms after the target signal), which reflect, respectively, the physiological processes associated with the preparation and execution of the movement. A numerical value of the ERP amplitude was obtained for each participant separately for each time window and for each EEG channel. Since ERP in different parts of the head can deviate from the zero value of the potential both up (positive peak) and down (negative peak), then the numerical values of the amplitude can be both positive and negative. Thus, our data takes into account both spatial (the name of the channel, i. e. its position on the head) and temporal (the first or second ERP window) characteristics of the brain response to the task in the stop-signal paradigm, as well as the electrical direction of the reaction (positive or negative peak amplitude values).

For each examined individual, the data dimension was 2×64 values. Since 50 participants were included in each group of people, the data size for each of our samples is approximately 50×2×64, and the total size of the data set is 100×2×64.

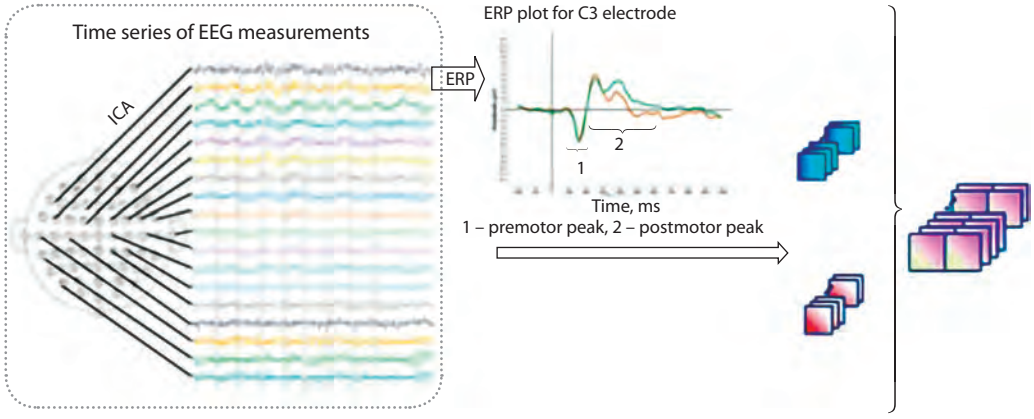


Fig. 1. The scheme of obtaining input data for the neural network.

Designing the structure and framework of a neural network

Since the input set of ERP data is small, a non-deep neural network was designed to predict whether an individual participated in long-term meditations or not. However, the initial EEG recording also has time series characteristics, so a convolutional neural network was additionally used for its analysis as a deep neural network for training and prediction. The main components of the convolutional neural network include convolutional layers, pooling layers, and fully connected layers.

In our case, the input layer of the convolutional network receives EEG data transformed into a two-dimensional matrix with a sample size of 2×64, where each row represents an individual ERP peak and each column represents an EEG recording channel. The hidden layer of the convolutional neural network includes three common architectures: a convolutional layer, a pooling layer, and a fully connected layer. We used the Conv1d() tool in PyTorch as the convolutional kernel, which prevented overfitting caused by using more complex convolutional kernels with more parameters (https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html#torch.nn.Conv1d, 21.02.2023).

The parameters of the convolutional layer include the kernel size, stride size, and padding, which collectively determine the size of the output feature map of the convolutional layer and are hyperparameters of the convolutional neural network. Due to the characteristics of EEG data, there are both spatial and temporal relationships, so we developed two schemes. The first scheme involves using a total of two one-dimensional convolutions. One extracts spatial features, which represent connections between ERP peaks in different electrode channels, and the other extracts temporal features. In this scheme, the PyTorch Conv1d() function wrapper was used to complete the corresponding function. The second approach involves applying only one one-dimensional convolution, but this convolution can extract both temporal and spatial features, for which the PyTorch Conv1d() function wrapper was also chosen.

The convolutional layers contain activation functions that help represent complex objects. In our study, three activation functions were used: sigmoid(), relu(), and softmax() from PyTorch (https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html, 15.04.2023). After extracting objects in the convolutional layer, the output feature map was passed to the pooling layer for object selection and information filtering. The pooling layer selects the pooling region in the same way as the kernel scanning stage of the convolutional layer, which is controlled by the pooling size, stride size, and padding. The convolutional and pooling layers in the convolutional neural network can extract features from the input data. The role of the fully connected layer is to nonlinearly combine the extracted features to obtain output data. In our case, two fully connected layers were created to prevent overfitting due to the small size of the dataset, for which the Linear() tool in PyTorch was applied. A fully connected layer is typically located before the output layer in a convolutional neural network. We used different loss and activation functions during training based on these two scenarios to improve the accuracy and performance of the model.

According to the above-described scheme, four network structures were designed and used for classifying surveyed individuals (Fig. 2). The only difference between these four architectures lies in the number of convolutional layers and the number of output neurons at the end.

In *the first structure*, a convolutional layer is used to extract both temporal and spatial objects. Then, two fully connected layers are used, and two values are output after normalization using the softmax activation function. Cross-entropy is used as the loss function, and Adam is used as the gradient descent algorithm.

The second structure also uses a convolutional layer to extract both temporal and spatial objects. Then, two fully connected layers are used, and the value is output after activation with the sigmoid function. Binary cross-entropy is used as the loss function, and Adam is used as the gradient descent algorithm.

The third structure uses two types of convolutions to extract spatial and temporal characteristics of the data, respectively. Then, two fully connected layers are used, and two values are output after normalization using the softmax activation function. Cross-entropy is used as the loss function, and Adam is used as the gradient descent algorithm.

Finally, *the fourth structure* uses two types of convolutions to extract spatial and temporal characteristics of the data, respectively. Then, two fully connected layers are used, and the value is output after activation with the sigmoid function. Binary cross-entropy is used as the loss function, and Adam is used as the gradient descent algorithm.

Optimal hyperparameters were found for each structure and are described in the model evaluation section.

Neural network training

The process of training an artificial neural network can be divided into four stages: initialization, forward propagation, backward propagation, and weight update.

During initialization, we assigned random initial values to each parameter (weights and biases) of the neural network to break symmetry and allow each neuron to have a different gradient and learn different functions. Later, during hyperparameter search, we determined the optimal initialization function for each architecture. During forward propagation, the training data (input and output) were fed into the neural network, and the activation value of each neuron was calculated sequentially from the input layer to the hidden layer, and then to the output layer according to the structure of the neural network. The activation values were obtained from the linear combination of the input data and weights plus bias, followed by a non-linear function such as sigmoid or ReLU. The goal of forward propagation was to obtain the predicted result of the neural network and compare it with the true result. The goal of backward propagation was to obtain the gradient of each parameter, which can be used to update the parameters. In our case, we used cross-entropy loss function and binary cross-entropy loss function for this purpose (https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html, 20.03.2023). The cross-entropy loss function was used to measure the distance between the probability distribution predicted by the model and the true probability distribution. Using this, we evaluated the performance of the model and

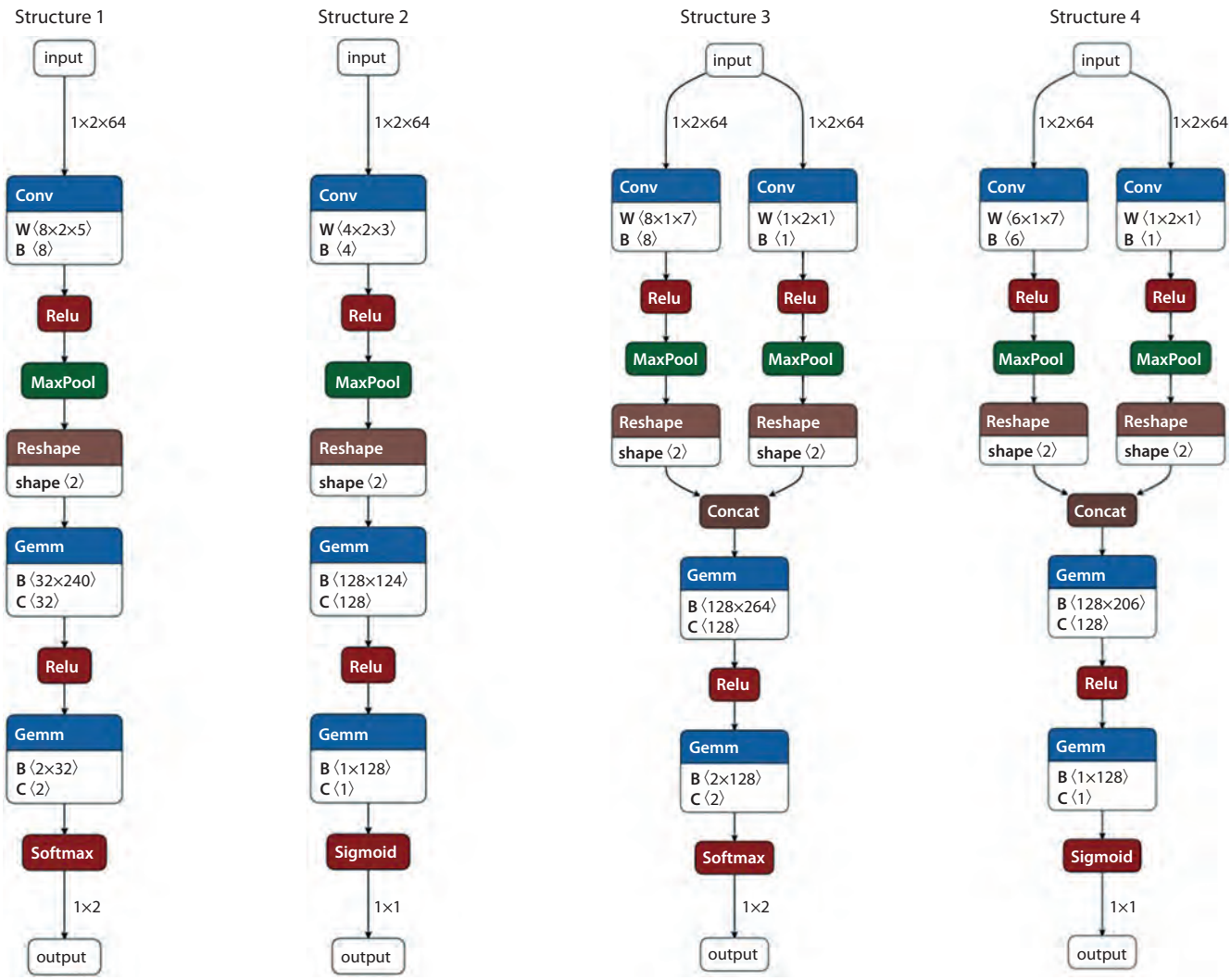


Fig. 2. Flowcharts of four models (structures) for the neural network architecture.

selected the optimal model and parameter by comparing the loss values of different models or different parameters.

Each parameter is updated with a certain learning rate (step size) according to its gradient, so that the loss function decreases. The goal of weight update is to optimize the parameters of the neural network so that it can better fit the training data. For this task, we applied the Adam optimization method. Adam is an algorithm for stochastic gradient descent with adaptive momentum, which was proposed at the ICLR conference in 2015 and has become one of the most popular and effective optimizers in deep learning. Adam combines two classical optimization algorithms, Adagrad and RMSProp, which are capable of handling sparse gradients and non-stationary objective functions, and uses the idea of momentum to accelerate convergence. Adam is equivalent to having a separate learning rate for each parameter, and this learning rate is adaptively adjusted according to the change in gradient. Specifically, when the gradient is large, the estimate of the second moment increases, which reduces the learning rate. When the gradient is small or sparse, the estimate of the first moment increases, which increases the learning rate. This effectively avoids oscillations caused by a too large learning rate, or increased complexity of convergence caused by a too small learning rate, or even getting trapped in a local minimum or saddle point.

To reduce overfitting and better train the model, we used batch normalization. Batch normalization is an approach that solves the problem of vanishing gradients by improving the smoothing of losses, speeding up network convergence, and increasing accuracy (Loffe, Szegedy, 2015). This method normalizes the data in mini-batches so that the mean value is 0 and the standard deviation is 1. At the same time, two trainable parameters, scale and shift, are introduced so that the model can learn its corresponding distribution during backward propagation. To implement this function, we used the BatchNorm1d() tool from PyTorch.

Overfitting is a common problem in the process of training an artificial neural network, where the model performs well on the training set but poorly on the test set or new data, indicating poor generalization. In our case, the problem was in overfitting due to a small dataset. To solve this problem, we applied initialization, L2 regularization, and dropout, as

well as cross-validation to evaluate the model and select hyperparameters that best train the model, reducing overfitting to some extent. We used L2 regularization (weight decay), which involves adding a penalty term to the loss function proportional to the sum of squares of the model's parameters. L2 regularization can cause the model's parameters to tend towards smaller values, thereby reducing the model's sensitivity to noise or outliers. Random deactivation (dropout) means the random zeroing of certain neurons or connection layers with a certain probability during training, which reduces the number of model parameters, thereby increasing the reliability and generalization ability of the model.

Cross-validation is the reuse of data, splitting the resulting dataset, combination into various training and test sets, a training set for training the model and a test set for evaluating the quality of model prediction. We used the K-fold multiplication method as a cross-validation method to reduce overfitting.

Evaluation of model performance on training data

In accordance with the characteristics of the EEG data sample and the indicators of the benchmark classification model, we used the metrics "F1-score", "AUC" (area under curve), and "accuracy" as evaluation indicators for the model (https://keras.io/api/models/model_training_apis). The higher these indicators, the better the model's performance. F1-score and AUC are comprehensive evaluation indicators for classification models, but they have different inaccuracies. AUC is less affected by the ratio of positive and negative samples in the dataset. For the purposes of this development, it became clear that predicting a person with a high level of stress as a person with a low level of stress would mean fundamentally incorrect results. Therefore, we chose F1-score as the most prioritized indicator for evaluating the model's effectiveness. We evaluated the model's hyperparameters using five-fold cross-validation to select the most suitable hyperparameters to prevent overfitting and improve model performance.

The results of evaluating the model on the training dataset are presented in Figure 3. Looking at each of the selected indicators, we can see that model 2 showed the most effective classification. Its effectiveness exceeded 80 % for all selected indicators. Models 1 and 4 also show good classification results, while model 3 performs the worst. Therefore, we assume that the output of one neuron surpasses the use of two neurons in the EEG binary classification task. Binary cross-entropy loss is obviously more suitable for our classification task based on the available dataset. When evaluating the model's effectiveness, the number of samples was 100, with 51 individuals practicing meditation (low stress level) and 49 individuals not practicing meditation. The number of samples is balanced, so it does not significantly affect the training and performance of the model. Moreover, for data with only two ERP peaks in 64 electrode channels, one convolution extracting both temporal and spatial characteristics worked better than two convolutions extracting temporal and spatial characteristics separately.

Evaluation of model performance on independent data. To evaluate the performance of the model on independent data, we prepared EEG data obtained from 25 individuals who were not included in the training set. Out of these 25 individuals, 12 practiced meditation, while 13 did not. The equipment, experimental design, and preprocessing of the EEG data were the same as in the training set. In this part of the study, all previously trained models were tested on new data that was not included in the training set. Accuracy, reliability, responsiveness, F1-score, ROC-AUC, specificity, and sensitivity were used as performance indicators for evaluating the models. Despite using parameter initialization functions, the weights were still randomly initialized within a certain range. Therefore, we adjusted the initial value of the random number to ensure the stability of the model's performance.

The performance metrics for different models on the independent test set are shown in Figure 4. According to the test results, structure 4 showed the best results for most selected parameters. Structure 2 also achieved good results. This structure exhibited the lowest sensitivity to overfitting, indicating its higher reliability compared to structure 4.

Conclusion

In our study, a neural network was successfully developed that classifies individuals into groups practicing or not practicing meditation based on the analysis of their EEG data with an accuracy of approximately 80–85 %. We used an EEG dataset collected and collated during our own experiments, selecting the amplitude of the ERP peak before button press at 250–350 ms and the amplitude value of the peak after button press at 550–900 ms for 64 recording channels. The sample size was 1×2×64.

Four architectures of non-deep convolutional networks were developed, among which structures 2 and 4 performed best in tests on independent data samples. Structure 2, which

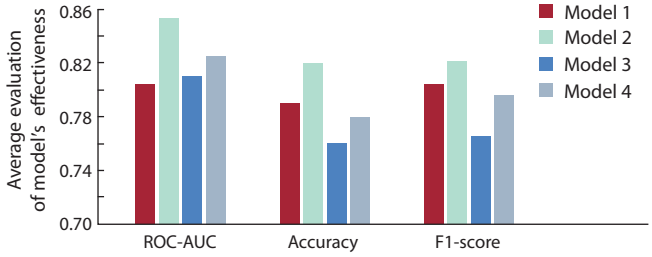


Fig. 3. Results of testing four different neural network models on the training sample.

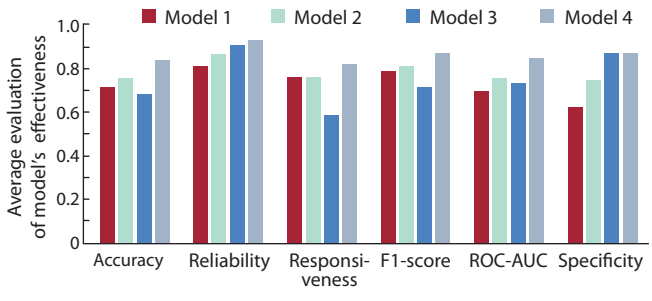


Fig. 4. Results of testing four different neural network models on the independent sample.

used a one-dimensional convolutional layer, pooling layer, and a two-layer fully connected network, showed the highest reliability. During the development of this model, it was noted that it was often prone to overfitting due to the limitation of the dataset size. This was mitigated by modifying the structure and scale of the model, specific network initialization parameters, regularization, random deactivation (dropout), and hyperparameter screening of cross-validation.

Overall, the approach proposed by us was tested on two relatively small samples of non-clinical subjects. A similar method on experimental data from the stop-signal paradigm had been previously tested by us in classifying samples of clinical patients with depressive disorders and healthy individuals (Zelenskih et al., 2022). The results of the research presented in this article complement the previous work, as they demonstrate that despite the small sample sizes, the convolutional neural network method allows to achieve a high level of accuracy in classifying different independent groups of people differing in stress levels. Taken together, the results of both studies show that applying neural networks to data obtained from individuals during the stop-signal paradigm is a promising method for assessing their stress levels and the severity of anxiety-depressive symptoms. It should be noted that the results of M.O. Zelenskih and colleagues' study are based solely on the application of behavioral data obtained in the stop-signal paradigm. The results of our new publication are based on the analysis of brain electrical responses obtained in the same experiment. The continuation of our research should involve the application of convolutional neural networks for the simultaneous analysis of behavioral and neurobiological data in order to more accurately classify participants based on their stress levels.

It is important to note that most standard methods for assessing stress levels or predisposition to anxiety-depressive disorders are based on the use of psychological questionnaires or interviews with a psychiatrist (e. g., Beck et al., 1988). However, such methods have a disadvantage: patients may not want to inform the interviewer about their condition or may inaccurately assess themselves. Inaccurate self-assessment by the patient is often the cause of incorrect conclusions regarding their susceptibility to illness (Nock et al., 2010). Another approach is based on the analysis of behavioral or neurophysiological reactions to emotional stimuli. Such stimuli can be either photographs of faces expressing the patient's or other people's emotional states (Quevedo et al., 2016), or emotional messages (Bocharov et al., 2020). This method allows for an objective assessment of the degree of impairment of the brain's affective functions but is less sensitive to changes in a person's overall ability to self-control behavior. Our proposed method, on the other hand, is based on the use of non-emotional stimuli to induce a complex sensorimotor reaction that requires either activation or inhibition of movement. Our approach allows for the assessment of the overall level of self-control of behavior but does not provide an opportunity to assess the patient's affective state. It is obvious that these three approaches (i. e., testing using questionnaires, analysis of reactions to affective stimulation, and analysis of reactions in motor control tasks) are mutually complementary, i. e., they should all be used together for a more detailed assessment of the same patient. Although our proposed approach currently requires further testing, it may yield significant results in the future in the development of diagnostic tools for stress-induced diseases.

References

Aftanas L., Golosheykin S. Impact of regular meditation practice on EEG activity at rest and during evoked negative emotions. *Int. J. Neurosci*. 2005;115(6):893-909. DOI 10.1080/00207450590897969

Atchley R., Klee D., Memmott T., Goodrich E., Wahbeh H., Oken B. Event-related potentials correlates of mindfulness meditation competence. *Neuroscience*. 2016;320:83-92. DOI 10.1016/j.neuroscience.2016.01.051

Band G.P.H., van der Molen M.W., Logan G.D. Horse-race model simulations of the stop-signal procedure. *Acta Psychol*. 2003;112(2):105-142. DOI 10.1016/s0001-6918(02)00079-3

Beck A.T., Steer R.A., Garbin M.G. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clin. Psychol. Rev.* 1988;8(1):77-100. DOI 10.1016/0272-7358(88)90050-5

Bocharov A.V., Savostyanov A.N., Tamozhnikov S.S., Merkulova E.A., Saprigyn A.E., Proshina E.A., Knyazev G.G. Oscillatory dynamics of perception of emotional sentences in healthy subjects with different severity of depressive symptoms. *Neurosci. Lett*. 2020;728:134888. DOI 10.1016/j.neulet.2020.134888

Chiesa A., Calati R., Serretti A. Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *Clin. Psychol. Rev*. 2011;31(3):449-464. DOI 10.1016/j.cpr.2010.11.003

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*. 2004;134(1):9-21. DOI 10.1016/j.jneumeth.2003.10.009.

Hsieh M.T., Lu H., Lin C.I., Sun T.H., Chen Y.R., Cheng C.H. Effects of trait anxiety on error processing and post-error adjustments: an event-related potential study with stop-signal task. *Front. Hum. Neurosci*. 2021;15:650838. DOI 10.3389/fnhum.2021.650838

Iwata N., Higuchi H.R. Responses of Japanese and American university students to the STAI items that assess the presence or absence of anxiety. *J. Pers. Assess*. 2000;74(1):48-62. DOI 10.1207/S15327752JPA740104

Khosla A., Khandnor P., Chand T. A comparative analysis of signal processing and classification methods for different applications based on EEG signals. *Biocybern. Biomed. Eng*. 2020;40(2):649-690. DOI 10.1016/j.bbe.2020.02.002

Kuh D., Ben-Shlomo Y., Lynch J., Hallqvist J., Power C. Life course epidemiology. *J. Epidemiol. Community Health.* 2003;57(10):778-783. DOI 10.1136/jech.57.10.778

Kuznetsova V.B., Knyazev G.G., Dorosheva E.A., Bocharov A.V., Savostyanov A.N. A role of personality and stress in the development of depressive symptoms in students. *Zhurnal Nevrologii i Psikhiatrii = Journal of Neurology and Psychiatry*. 2016;116(12):114-118. DOI 10.17116/jnevro2016116121114-118 (in Russian)

Loffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv.* 2015;1502.03167. DOI 10.48550/arXiv.1502.03167

Logan G.D., Cowan W.B. On the ability to inhibit thought and action: a theory of an act of control. *Psychol. Rev*. 1984;91(3):295-327. DOI 10.1037/0033-295X.91.3.295

McCrae R.R., Costa P.T., Jr., Ostendorf F., Angleitner A., Hrebícková M., Avia M.D., Sanz J., Sánchez-Bernardos M.L., Kusdil M.E., Woodfield R., Saunders P.R., Smith P.B. Nature over nurture: temperament, personality, and life span development. *J. Pers. Soc. Psychol*. 2000;78(1):173-186. DOI 10.1037//0022-3514.78.1.173

Nock M.K., Park J.M., Finn C.T., Deliberto T.L., Dour H.J., Banaji M.R. Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychol. Sci*. 2010;21(4):511-517. DOI 10.1177/0956797610364762

Quevedo K., Scott R.N.H., Martin J., Smyda G., Keener M., Oppenheimer C.W. The neurobiology of self-face recognition in depressed adolescents with low or high suicidality. *J. Abnorm. Psychol*. 2016; 125(8):1185-1200. DOI 10.1037/abn0000200

Saeed S.A., Cunningham K., Bloch R.M. Depression and anxiety disorders: benefits of exercise, yoga, and meditation. *Am. Fam. Physician*. 2019;99(10):620-627

Savostyanov A.N., Tsai A.C., Liou M., Levin A.E., Lee J.D., Yurganov A.V., Knyazev G.G. EEG-correlates of trait anxiety in the stop-signal paradigm. *Neurosci. Lett*. 2009;449(2):112-116. DOI 10.1016/j.neulet.2008.10.084

Savostyanov A.N., Tamozhnikov S.S., Bocharov A.V., Saprygin A.E., Matushkin Y., Lashin S., Kolpakova G., Sudobin K., Knyazev G. The effect of meditation on comprehension of statement about oneself and others: a pilot ERP and behavioral study. *Front. Hum. Neurosci*. 2020;13:437. DOI 10.3389/fnhum.2019.00437

Zelenskih M.O., Saprygin A.E., Tamozhnikov S.S., Rudych P.D., Lebedkin D.A., Savostyanov A.N. Development of a neural network for diagnosing the risk of depression according to the experimental data of the stop signal paradigm. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):773-779. DOI 10.18699/VJGB-22-93

ORCID ID

A.E. Saprygin orcid.org/0000-0001-6789-2953
A.N. Savostyanov orcid.org/0000-0002-3514-2901

Acknowledgements. The development and testing of the neural network and the collection of EEG data from meditators was carried out as part of the budget project of the ICG SB RAS No. FWNR-2022-0020. The collection of EEG data from non-meditators, as well as the preprocessing of all EEG data, was carried out as part of the project of the Russian Scientific Foundation No. 22-15-00142 "fMRI and EEG correlates of the focus of attention on oneself as a factor of propensity to affective disorders".

Conflict of interest. The authors declare no conflict of interest.

Received July 11, 2023. Revised September 10, 2023. Accepted September 13, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):859-868
DOI 10.18699/VJGB-23-99

Original Russian text https://vavilovj-icg.ru/

Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods

E.G. Komyshev[1], M.A. Genaev[1, 2, 3], I.D. Busov[1, 3], M.V. Kozhekin[2], N.V. Artemenko[2, 3], A.Y. Glagoleva[1], V.S. Koval[1], D.A. Afonnikov[1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
komyshev@bionet.nsc.ru

Abstract. The pigment composition of plant seed coat affects important properties such as resistance to pathogens, pre-harvest sprouting, and mechanical hardness. The dark color of barley (*Hordeum vulgare* L.) grain can be attributed to the synthesis and accumulation of two groups of pigments. Blue and purple grain color is associated with the biosynthesis of anthocyanins. Gray and black grain color is caused by melanin. These pigments may accumulate in the grain shells both individually and together. Therefore, it is difficult to visually distinguish which pigments are responsible for the dark color of the grain. Chemical methods are used to accurately determine the presence/absence of pigments; however, they are expensive and labor-intensive. Therefore, the development of a new method for quickly assessing the presence of pigments in the grain would help in investigating the mechanisms of genetic control of the pigment composition of barley grains. In this work, we developed a method for assessing the presence or absence of anthocyanins and melanin in the barley grain shell based on digital image analysis using computer vision and machine learning algorithms. A protocol was developed to obtain digital RGB images of barley grains. Using this protocol, a total of 972 images were acquired for 108 barley accessions. Seed coat from these accessions may contain anthocyanins, melanins, or pigments of both types. Chemical methods were used to accurately determine the pigment content of the grains. Four models based on computer vision techniques and convolutional neural networks of different architectures were developed to predict grain pigment composition from images. The U-Net network model based on the EfficientNetB0 topology showed the best performance in the holdout set (the value of the "accuracy" parameter was 0.821).
Key words: digital image analysis; machine learning; barley grains; pigment composition.

For citation: Komyshev E.G., Genaev M.A., Busov I.D., Kozhekin M.V., Artemenko N.V., Glagoleva A.Y., Koval V.S., Afonnikov D.A. Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):859-868. DOI 10.18699/VJGB-23-99

Определение содержания меланина и антоцианов в зернах ячменя на основе анализа цифровых изображений методами машинного обучения

Е.Г. Комышев[1], М.А. Генаев[1, 2, 3], И.Д. Бусов[1, 3], М.В. Кожекин[2], Н.В. Артеменко[2, 3], А.Ю. Глаголева[1], В.С. Коваль[1], Д.А. Афонников[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
komyshev@bionet.nsc.ru

Аннотация. Пигментный состав оболочек семян растений влияет на такие важные их свойства, как устойчивость к действию патогенов, прорастание на корню, а также механическая прочность. У ячменя (*Hordeum vulgare* L.) темная окраска зерен может быть обусловлена синтезом и накоплением двух групп пигментов. Голубая и фиолетовая окраска зерна связана с синтезом антоцианов. Серую и черную окраску придают пигменты меланины. Данные пигменты могут накапливаться в оболочках зерна независимо либо совместно, поэтому визуально определить, накопление каких именно пигментов придает темный цвет зерна, затруднительно. Для точного определения наличия/отсутствия пигментов используются химические и генетические методы, которые дороги и трудоемки. Поэтому создание нового метода для быстрой оценки наличия определенных пигментов в зерновке является актуальной задачей, решение которой поможет при исследовании механиз-

мов генетического контроля пигментного состава зерна. Настоящая работа посвящена разработке метода оценки пигментного состава зерен ячменя на основе анализа цифровых изображений с помощью алгоритмов компьютерного зрения и машинного обучения. Разработан протокол съемки для получения двумерных цифровых цветных изображений зерен. С использованием данного протокола получено 972 изображения для 108 образцов ячменя. Каждый образец мог содержать пигменты антоцианы и/или меланины. Для точного определения содержания пигментного состава образцов применялись химические методы. Для предсказания пигментного состава зерна на основе изображений было разработано четыре модели, основанных на методах компьютерного зрения и сверточных нейронных сетях различной архитектуры. Лучшую производительность на отложенной выборке показала модель сети U-Net, основанная на топологии EfficientNetB0 (значение параметра «точность» составило 0.821).
Ключевые слова: анализ цифровых изображений; машинное обучение; зерна ячменя; пигментный состав.

Introduction

The color of cereal grain shell is an important trait characterizing the pigments and metabolites contained in it. The presence of pigments in the shell affects various technological properties of the grain (Souza, Marcos-Filho, 2001; Flintham et al., 2002). Grains with dark grain coloration are more cold- and drought-tolerant and also have increased resistance to pathogens (Ceccarelli et al., 1987; Choo et al., 2005). Such properties of colored grains are associated with high antioxidant content as well as additional mechanical hardness of grain shells (Ferdinando et al., 2012; Jana, Mukherjee, 2014). The dark color of barley grains occurs due to the synthesis and accumulation of two groups of pigments. Blue and purple coloration of the grain shell is associated with the biosynthesis of anthocyanins. Gray and black color of barley grains is caused by melanin pigment. These two types of pigments can accumulate in the grain shell depending on the genotype both individually and together. Therefore, it is difficult to determine which pigments cause dark grain color by eye.

A number of regulatory genes and genes encoding enzymes involved in pigment biosynthesis control grain shell coloration. Currently, the pathway of anthocyanin biosynthesis has been investigated quite well, but the molecular mechanisms of melanin biosynthesis are still poorly understood (Shoeva et al., 2018; Glagoleva et al., 2020). When studying the mechanisms of genetic control of grain coloration, breeders and geneticists need to assess the pigment content of grain shells. Chemical methods for estimating pigment content allowed to accurately determine the presence/absence of pigments; however, they are expensive and labor-intensive. Other approaches to solving this problem include spectrophotometers, spectrometers, and hyperspectral cameras. However, these cameras are expensive, especially those with high resolution, both spatial and spectral. An alternative is the use of digital RGB cameras that produce high-quality images with high spatial and color resolution (Afonnikov et al., 2016; Li et al., 2020; Kolhar, Jagtap, 2023). In this regard, methods for estimating color and textural characteristics of cereal grains based on the analysis of two-dimensional images acquired by digital cameras or scanners have recently been intensively developed in the field of grain phenotyping (Komyshev et al., 2020; Sharma et al., 2021; Afonnikov et al., 2022; Arif et al., 2022; Khojastehnazhand, Roostaei, 2022; Wang, Su, 2022).

The aim of this work is to develop a method for estimating the pigment composition of barley grain based on the analysis of digital images using computer vision and machine learning algorithms.

Materials and methods

Plant material. Grains of 39 barley accessions with dark colored grain and 40 accessions with light grains were selected for the study. The material was obtained from the barley collection of the All-Russian Institute of Plant Genetic Resources named after N.I. Vavilov (VIR, https://www.vir.nw.ru), the barley collection of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (ICG, https://www.icgbio.ru) and the material from the Oregon Wolfe Barleys population (OWB, https://barleyworld.org/owb). The material description is summarized in Supplementary Material 1[1]. Twenty-nine barley accessions from the VIR collection with different combinations of pigments in the grain were also separately selected (Supplementary Material 2). The material included hulled and hulless barley accessions. 58 hulled and 21 hulless accessions were chosen to create training and test datasets. 22 hulled and 7 hulless accessions were used in the holdout dataset.

Chemical methods for determining the pigment composition of grains. To determine the presence of anthocyanins in the grain shell, extraction in 1 % HCl solution in methanol, followed by detection of pink coloration of the solution, was performed (Abdel-Aal, Hucl, 1999). The presence of melanin was determined using 2 % NaOH, in which melanin extraction occurs and stains the solution dark (Downie et al., 2003). Based on this method, each of the accessions was assigned a type of pigmentation based on the presence of these pigments ("anthocyanins", "melanins") or "no pigments" if both pigments were absent in the grain shell. The presence of pigments of a particular type in the accession seed shells is summarized in Supplementary Materials 1 and 2.

Image acquisition. Color images of grains were obtained using a Canon EOS 600D digital camera, Canon EF 100mm f/2.8 Macro USM lens with a resolution of 18 MP. A 55 mm diameter plastic Petri dish filled with grains without gaps was placed on a white A3 sheet of matte paper. Diffusing light was placed on the sides, and the camera was fixed on a tripod from above, with the lens vertically downward (Supplementary

[1] Supplementary Materials 1–8 are available at:
https://vavilov.elpub.ru/jour/manager/files/Suppl_Komyshev_Engl_27_7.pdf

Fig. 1. A typical image obtained by the protocol for barley grain phenotyping.

Material 3). Images were saved in JPEG format. Figure 1 shows an example of an image resulting from the protocol.

The Petri dish contained about 100–160 grains. For each accession, 9 images of its replicas were obtained by randomly mixing grains in a Petri dish.

Data markup. In order to develop a segmentation algorithm for 212 images of 59 randomly selected accessions, manual marking of grains and Petri dish boundaries was performed using the LabelMe program (https://github.com/wkentaro/labelme). An example of a labeled image fragment is shown in Supplementary Material 4. In addition, each image was labeled according to the pigmentation type of the corresponding accession based on experimentally obtained data.

Prediction of grain pigmentation based on machine learning methods. The general scheme for pigmentation type prediction involved segmenting the image into the background and the area occupied by grains and predicting the presence of pigments of a particular type using three methods: (1) a Random Forest algorithm using image color descriptors; (2) a convolutional neural network of the ResNet-18 architecture; and (3) a convolutional neural network of the EfficientNetB0 architecture.

Data partitioning scheme for validation and testing. For machine learning methods, the images were divided into three datasets: training (60 % of data: 423 images, 47 accessions); validation (20 % of data: 144 images, 16 accessions); and test (20 % of data: 144 images, 16 accessions). A holdout dataset of 29 accessions including 261 images was used for the final accuracy evaluation. Stratification was used to partition the acquired images (see Supplementary Material 5). Data on the partitioning of the accessions into subsamples are presented in Supplementary Material 5.

Evaluating the accuracy of grain image classification. The output of the trained classification models for each image was represented by two binary numbers, each of which characterized the presence or absence of anthocyanins and melanin. To evaluate the accuracy of the method on the test dataset for each image, the predicted set of such numbers and the true set were compared. The following metrics were calculated based on these comparisons: true positive class predictions (TP), true negative class predictions (TN), total number of positive (P) and negative (N) class representatives. Based on these values, the ACC (accuracy) was calculated according to the formula:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}.$$

A model for identifying the grain region in an image. To distinguish grains in Petri dishes from the background, the U-Net neural network model with a ResNet-18 encoder was used. The U-Net model was chosen as this architecture had been developed specifically for biomedical image segmentation (Ronneberger et al., 2015). The model is based on the use of convolution and consists of two parts: an encoder and a decoder (Fig. 2). The full-size image at the input of the network is transformed by the encoder through several steps including two consecutive convolution layers of size 3×3 followed by a ReLU transform (labeled as 'conv 3×3, ReLU' layers in Fig. 2) and pooling with a maximum 2×2 function with a step size of 2 (labeled as 'max pool 2×2' layers). The encoder performs downsampling of the image. The decoder, on the other hand, performs upsampling of the image using a series of inverse pooling operations that expand the feature map. This is followed by 2×2 convolution, which reduces the number of feature channels (labeled as 'up-conv 2×2' layers). This is followed by a concatenation with an appropriately edge-cropped feature map from the compressive path and two 3×3 convolutions (labeled as 'copy and crop' layers in Fig. 2), after each of which a ReLU operation is applied.

Segmentation allowed us to select a region of the Petri dish with grains in the image, which was used to calculate their color descriptors. For each image, 2,380 numerical parameters characterizing the pixel color of the grains were extracted. These are average values of channel intensities for 4 color spaces (RGB, HSV, Lab, YCrCb), values of histograms of color component intensity distributions, etc. Detailed description of the obtained characteristics is given in Supplementary Material 6.

Data filtering. We removed from the prediction input data features, the values of which were identical for all images or did not exceed the value of 0.01 for more than 20 % of images. Additionally, we selected features with pairwise Spearman correlation coefficient less than 0.97 in the image dataset to eliminate redundancy. As a result, 345 color features out of 2,380 remained for our analysis.

Data analysis. In order to estimate the distribution of accessions in the feature space under study, the principal component method (Jolliffe, 2002) and t-SNE algorithm for the nonlinear dimensionality reduction (van der Maaten, Hinton, 2008) were used. These methods allow visualization of multidimensional data by mapping objects from a multidimensional space to a lower dimensional space.

A model for classification of pigment composition of grains based on color descriptors by the Random Forest method

The classification of grain images into four classes was considered: (1) no pigmentation, (2) presence of anthocyanins only, (3) presence of melanin only, (4) presence of both

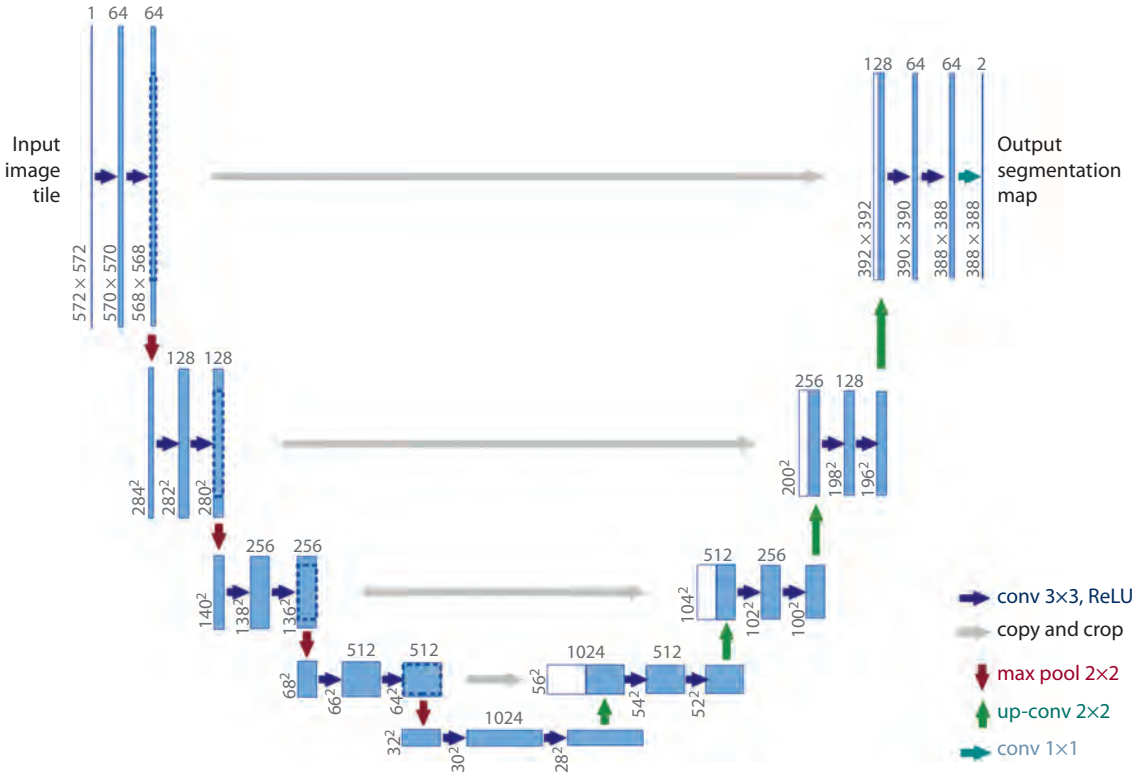


Fig. 2. U-Net network architecture used for image segmentation into grain and background regions, from (Ronneberger et al., 2015).

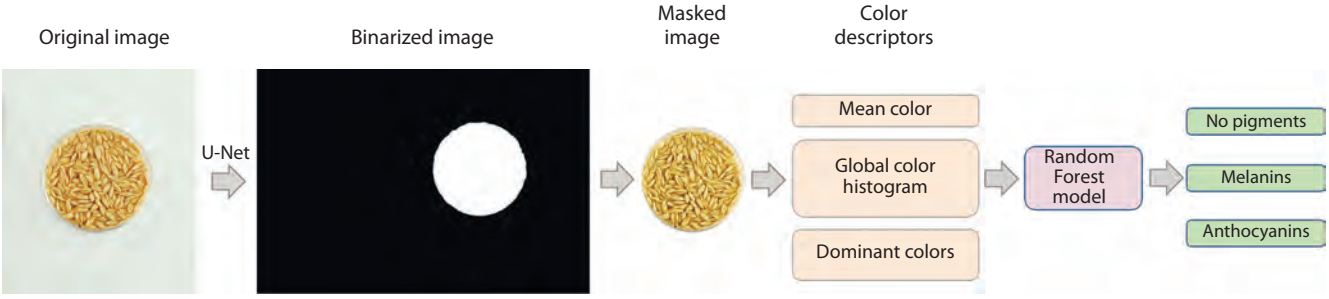


Fig. 3. Scheme of the barley grain classification model based on the Random Forest algorithm using color descriptors (the RF13 model).

anthocyanins and melanins. The first classification model was built using the Random Forest algorithm implemented in the Scikit-learn package (Pedregosa et al., 2011). The values of 345 color descriptors described above were used as input. The data processing scheme for this model is shown in Figure 3. Additionally, using the principal component method, the number of features was reduced to 13, which explain 81.2 % of the variance in the data and give the maximum accuracy on the test dataset. We have termed this classification model RF13.

Grain pigment composition classification models based on deep machine learning

ResNet-18 architecture network-based classification model. In addition to the above described RF13 model, three models based on deep machine learning methods were implemented to predict the grain shell pigmentation type. These methods are now widely used to analyze plant images and have been shown to be highly accurate.

One of the models is the ResNet-18 neural network architecture (He et al., 2016). ResNet is a family of convolutional neural networks (CNNs) of similar architecture differing in the number of layers (18, 34, 50, 101, and 152). In this work, we used a model with 18 layers as the simplest and fastest one. It consists of 17 layers in series including convolution transform, connected by an alternate path for the signal and one full-link layer (Fig. 4). Every four layers, a subsampling operation takes place, where the length and width of the layer becomes 2 times smaller and the number of channels doubles. In Figure 4, these are the layers labeled as "3×3 convolution, N", where N is the number of channels.

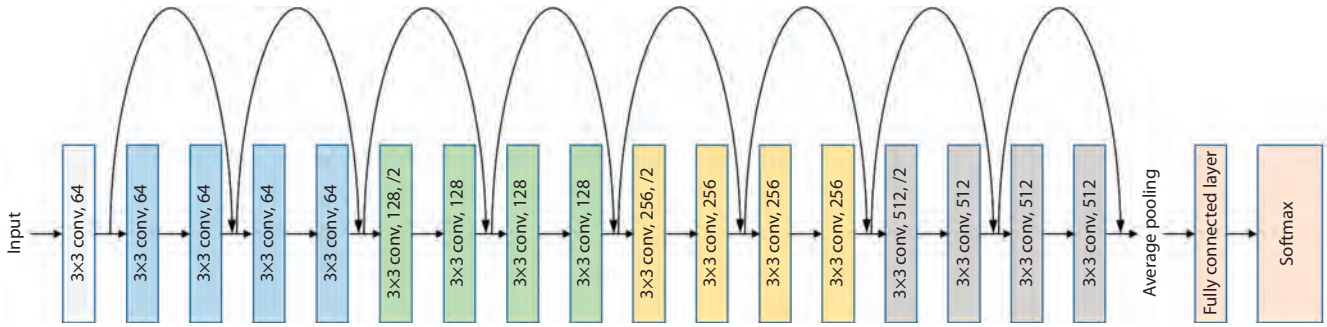


Fig. 4. Schematic diagram of ResNet-18 network architecture.
Different-colored rectangles show network layers of different structure.

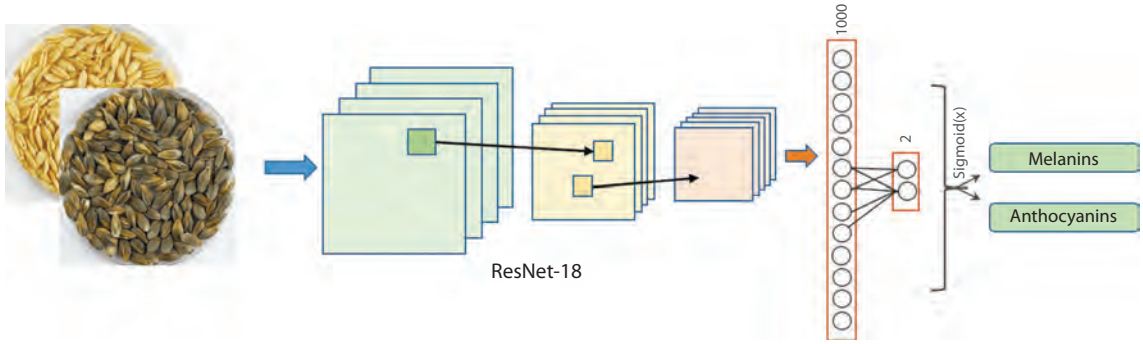


Fig. 5. Schematic of the ResNet-18 model of barley grain image classification based on convolutional neural network.

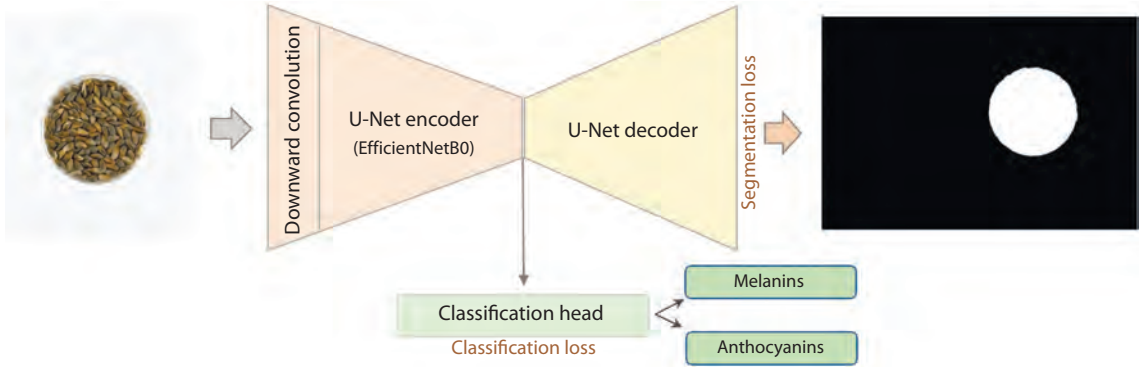


Fig. 6. Schematic of the U-Net+ClassHead model based on U-Net segmentation with a head for simultaneous segmentation and classification of barley grain images by the presence/absence of anthocyanins or melanin.

The input of the network was rectangular images, which included regions of Petri dishes (Fig. 5). The output layer included two numbers between 0 and 1 predicting the presence (1) of melanin or anthocyanins. In case the number value was greater than 0.5, the corresponding pigment was considered to be present in the grain shell. This method allowed us to classify images based on the presence of the two pigments in the grains both individually and jointly, and to identify their absence in case both numbers were less than 0.5. This classification model was termed ResNet-18 in our work.

A segmentation-based model with a head for classification. The neural network parameters that were obtained during image segmentation using the U-Net algorithm can be used to classify grains by the presence of pigments. This allows to improve the prediction accuracy for algorithms and to solve two problems simultaneously (segmentation and classification). To this end, an additional output classification layer ("classification head") was added to the existing segmentation-based model with U-Net architecture (Fig. 6). The output of this layer, as in the ResNet-18 model, contains two numbers to determine the presence of anthocyanins and/or melanin in

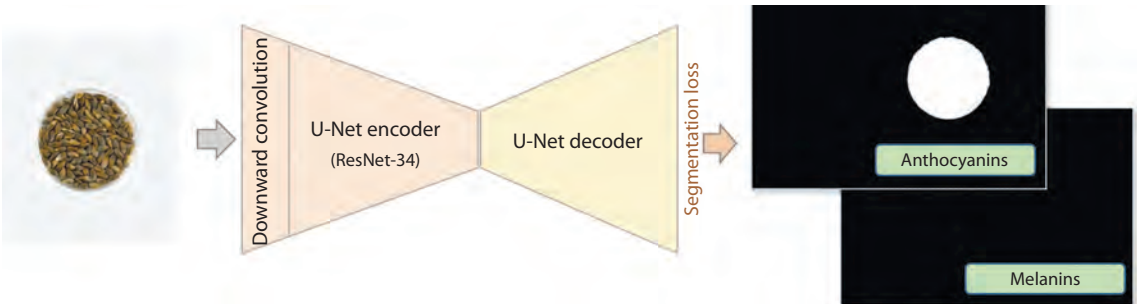


Fig. 7. Schematic of the U-Net+ClassSegment model for classification based on 2-channel segmentation of barley grain images by the presence/absence of anthocyanins or melanin.

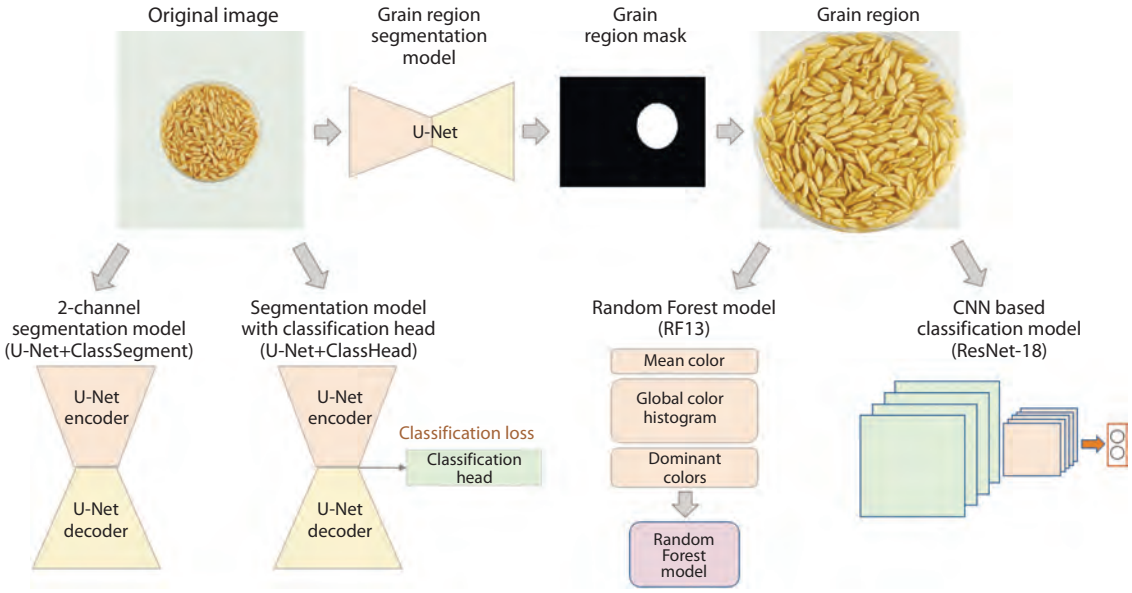


Fig. 8. General scheme of barley grain image analysis by the models proposed in this paper.

the grains (see Fig. 6). For this network, the coder topology of the EfficientNetB0 architecture was used (Tan, Le, 2019). This network topology allowed not only to segment the image by selecting a region of grains in a Petri dish on the image, but also to simultaneously perform classification of the whole image based on the presence or absence of the two pigments. This classification model was termed U-Net+ClassHead in the paper.

2-channel segmentation model. For image classification, a modified U-Net can be used to segment each pixel in the image based on the presence of a particular pigmentation. This network outputs a two-channel mask, in which each channel segments the image region if the grain shells contain a particular pigment (Fig. 7). This model, U-Net+ClassSegment, was based on the U-Net architecture with the ResNet-34 encoder. To determine the class of the whole image, we considered that if a single pixel was classified as containing a pigment after segmentation, the whole sample was considered to contain that pigment.

Other technical parameters of training models such as the number of training epochs, batched size, loss function used and optimizer parameters are given in Supplementary Material 7.

Thus, two classification models based on U-Net segmentation of the original image (U-Net+ClassHead and U-Net+ClassSegment) and two classification models for which the grain region in the original images was separately extracted using the U-Net segmentation model (RF13 and ResNet-18) were considered in this paper. The general scheme of image analysis by the proposed segmentation and classification models is shown in Figure 8.

Results

Color characteristics of grains

PCA and t-SNE methods were applied to map grain images for accessions into a generalized feature space of dimension 2 using 345 informative features (see Materials and methods). The feature values were subjected to normalization before this

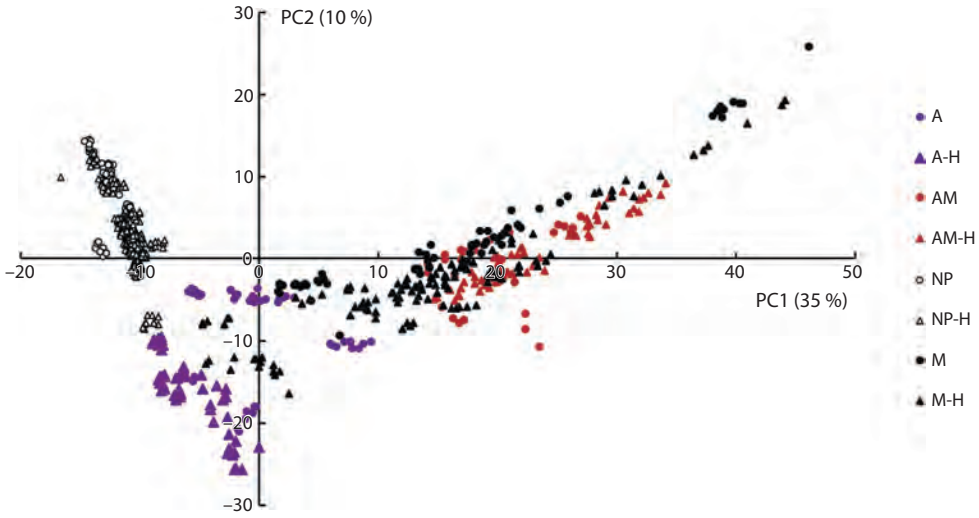


Fig. 9. Scattering diagram of the grain images for barley accessions in the space of the first two components derived from PCA for the color characteristics of grains.

The *X* axis is the PC1 component, the *Y* axis is the PC2 component. Fractions of dispersion for the components are given in parentheses. Grain type designations for pigments and hull presence are shown on the right (A, AM, M, NP – anthocyanins, anthocyanins and melanin, melanin, and no pigments, respectively; H – hulled grains).

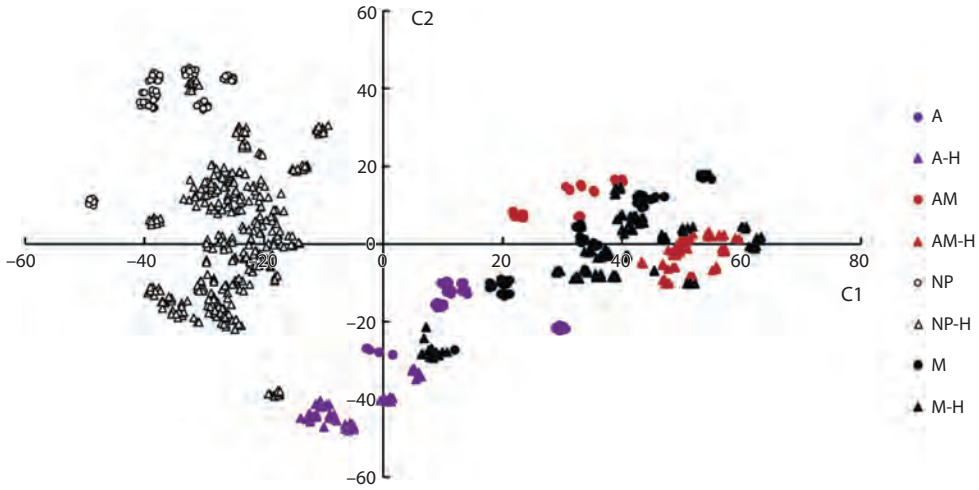


Fig. 10. Scatter diagram of the grain images for barley accessions in the space of the first two components resulting from the t-SNE algorithm for the color characteristics of grains.

The *X* axis is the C1 component, the *Y* axis is the C2 component. Fractions of dispersion for the components are given in parentheses. Grain type designations for pigments and hull presence are shown on the right (A, AM, M, NP – anthocyanins, anthocyanins and melanin, melanin, and no pigments, respectively; H – hulled grains).

analysis (to obtain mean equal to 0 and standard deviation equal to 1). Each point in PCA (Fig. 9) and t-SNE (Fig. 10) diagrams corresponds to a particular image.

These diagrams show that pigmented (filled markers) and non-pigmented (empty markers) grains are well separated in both diagrams (see Fig. 9 and 10). This separation is more pronounced in the t-SNE diagram (see Fig. 10). Images of grains with the presence of anthocyanins in the shell (purple icons) and those containing both pigments (red icons) are well separated. The areas occupied by these images in the diagrams do not overlap. At the same time, it is noticeable that the regions occupied by the images of grains with anthocyanins (filled purple markers) and melanin (filled black markers) overlapped. It is also clear from the diagrams that regions for the images of grains containing both anthocyanins and melanin and those containing only melanin have considerable overlap (the right part of the plots close to 0 values for the *Y* axis). Separating these two types of grains seems most problematic.

The influence the presence of the grain hull has on their color characteristics is also noticeable in the two graphs. First of all, the presence of the grain hull does not affect the separation of areas for different classes of grains by pigmentation except for the pair containing anthocyanins or melanin: hulled and hulless grains with the same type of pigmentation are closer to each other than grains with another type of pigmentation.

Table 1. Assessment of classification accuracy (ACC) of barley grain images based on anthocyanins and melanin presence in grain coat for four models on validation, test and holdout datasets

Classification model	Validation	Test	Holdout
RF13	0.896	0.903	0.652
ResNet-18	0.938	0.934	0.817
U-Net+ClassHead	0.906	0.962	**0.821**
U-Net+ClassSegment	0.917	0.903	0.819

Note. The best value for the holdout dataset is shown in bold.

Table 2. Parameters for evaluation of classification performance of barley grain images by anthocyanins and melanin presence in the grain coat for the U-Net+ClassHead model on test and holdout datasets

Parameter	Test		Holdout	
	Melanin	Anthocyanins	Melanin	Anthocyanins
F-measure	1.0	0.937	0.983	0.488
Sensitivity	1.0	0.881	1.0	0.389
Positive predictable value	1.0	1.0	0.966	0.656

This is particularly evident for grains without pigmentation (empty markers). For grains with melanin, one of the groups of hulled grains has color characteristics very similar to those of grains with anthocyanins presence (on the graphs, this group is located inside the area occupied by samples with anthocyanins and is far away from other grains containing melanin). At the same time, it is clearly visible that for grains of the same pigment class, hulled and hulless grains occupy different regions and are well separated (characteristic examples in Fig. 10: images of grains without pigmentation, images of grains with anthocyanins, and images of grains with anthocyanins and melanins). These results show that, in most cases, the presence of the hull does not affect separation by the type of grain pigmentation, but significantly affects the variation of shell color characteristics.

Classification of grains by pigment content

As a result of training the models to classify grain images by pigment content, accuracy estimates on validation, test and holdout datasets were obtained. They are presented in Table 1.

The best accuracy on the holdout dataset is achieved by the segmentation model with "classification head" (U-Net+ClassHead). The data on the parameters of performance estimates of this model are given in Table 2.

The prediction error matrix of the grain pigmentation type (Supplementary Material 8) allows us to determine that most of the model errors are in predicting the anthocyanin content of hulled grains, which is consistent with the PCA and t-SNE plots (see Fig. 9 and 10), where regions for hulled grain images containing melanin and anthocyanins overlap significantly with those containing only melanin. Moreover, the number of images with grain containing anthocyanins (A) predicted as not containing pigments (NP) is significantly larger than the number of images of grains without pigments (NP) predicted as anthocyanins (A). Errors are also observed for hulless grains, for which the presence of anthocyanins was erroneously not predicted. A small number of images of grains with melanins were predicted as "no pigments", some images of grains containing anthocyanins were identified as containing melanins.

The results of the non-parametric Mann–Whitney test showed that the accuracy of anthocyanins presence prediction differs significantly (p-value = 0.004) for hulless and hulled grains. For melanin presence, the hull does not significantly affect the prediction performance.

The U-Net+ClassSegment method showed slightly lower accuracy. It can be concluded that models that simultaneously solve several different tasks (multi-task learning) have better generalization ability. Both models based on this approach significantly outperform both the method based on Random Forest and color descriptors (lowest accuracy) and the ResNet-18 classification. It is worth noting that the accuracy results on the holdout dataset are lower than on the test dataset.

Discussion

Methods for analyzing digital RGB images to study the physiological properties of grains have been widely applied to cereals (Neuman et al., 1989; Huang et al., 2015; Sabanci et al., 2017; Kozłowski et al., 2019; Komyshev et al., 2020; Zykin et al., 2020). In particular, they are used to classify grains both by pigment composition and by variety.

In our work, we analyzed methods for classifying grains by color characteristics into classes based on the presence of two types of pigments. We showed that deep machine learning methods yield higher accuracy in grain classification than using color descriptors. Similar findings were obtained when classifying barley grains into species (Kozłowski et al., 2019). Our results also show that using a multi-task learning approach produces more accurate classification results.

The results on the holdout image dataset showed lower accuracy compared to the test dataset. Presumably, one of the reasons for this could be that the balance of labels of different classes in the training, validation, and test datasets was the same and was not close to the ratio in the holdout dataset. In particular, the number of images with grains without pigments in the holdout dataset was 1.5 times lower than in the training sample. For classification, such an image set appears to be the easiest case. Also, based on the extracted color descriptors, a binary classifier was trained that distinguished grains from the holdout dataset from other grains with ACC = 1. This implies that there are significant differences between these image series, which can be explained by the fact that grains from other collections were selected in the holdout dataset or the protocol for capturing these images was slightly different. This can explain the slight decrease in accuracy in the classification quality of the Random Forest model.

Our analysis also demonstrated that the presence of the hull affects grain color characteristics and, thus, the classification performance with respect to the pigment presence in the shell.

Conclusion

The proposed methods based on the analysis of digital images using computer vision and machine learning algorithms showed acceptable classification ability in the task of determining melanin and anthocyanins presence or absence in the barley grain shell. The results of this work showed that the application of the Random Forest algorithm based on color features is inferior to convolutional neural network approaches in the classification performance. This method proves to be sensitive to small changes in protocol or imaging conditions, losing generalization ability compared to convolutional neural networks. Possible ways to improve the model based on this algorithm are careful selection of features and preliminary normalization of the images fed to the input. The classical classification model architecture is inferior in accuracy to the 2-channel whole image segmentation model. Segmentation by U-Net neural network with "classification head", showed the best results (ACC = 0.821) and is the preferred choice in the task of determining the pigment content of barley.

References

Abdel-Aal E.S.M., Hucl P. A rapid method for quantifying total anthocyanins in blue aleurone and purple pericarp wheats. *Cereal Chem*. 1999;76(3):350-354. DOI 10.1094/CCHEM.1999.76.3.350

Afonnikov D.A., Genaev M.A., Doroshkov A.V., Komyshev E.G., Pshenichnikova T.A. Methods of high-throughput plant phenotyping for large-scale breeding and genetic experiments. *Russ. J. Genet*. 2016;52(7):688-701. DOI 10.1134/S1022795416070024

Afonnikov D.A., Komyshev E.G., Efimov V.M., Genaev M.A., Koval V.S., Gierke P.U., Börner A. Relationship between the characteristics of bread wheat grains, storage time and germination. *Plants*. 2022;11(1):35. DOI 10.3390/plants11010035

Arif M.A.R., Komyshev E.G., Genaev M.A., Koval V.S., Shmakov N.A., Börner A., Afonnikov D.A. QTL analysis for bread wheat seed size, shape and color characteristics estimated by digital image processing. *Plants*. 2022;11(16):2105. DOI 10.3390/plants11162105

Ceccarelli S., Grando S., Van Leur J.A.G. Genetic diversity in barley landraces from Syria and Jordan. *Euphytica*. 1987;36(2):389-405. DOI 10.1007/BF00041482

Choo T.M., Vigier B., Ho K.M., Ceccarelli S., Grando S., Franckowiak J.D. Comparison of black, purple, and yellow barleys. *Genet. Resour. Crop Evol*. 2005;52(2):121-126. DOI 10.1007/s10722-003-3086-4

Downie A.B., Zhang D., Dirk L.M.A., Thacker R.R., Pfeiffer J.A., Drake J.L., Levy A.A., Butterfield D.A., Buxton J.W., Snyder J.C. Communication between the maternal testa and the embryo and/or endosperm affect testa attributes in tomato. *Plant Physiol*. 2003; 133(1):145-160. DOI 10.1104/pp.103.022632

Ferdinando M.D., Brunetti C., Fini A., Tattini M. Flavonoids as antioxidants in plants under abiotic stresses. In: Ahmad P., Prasad M. (Eds.) Abiotic Stress Responses in Plants. New York: Springer, 2012;159-179. DOI 10.1007/978-1-4614-0634-1_9

Flintham J., Adlam R., Bassoi M., Holdsworth M., Gale M. Mapping genes for resistance to sprouting damage in wheat. *Euphytica*. 2002; 126:39-45. DOI 10.1023/A:1019632008244

Glagoleva A.Y., Shoeva O.Y., Khlestkina E.K. Melanin pigment in plants: current knowledge and future perspectives. *Front. Plant Sci*. 2020;11:770. DOI 10.3389/fpls.2020.00770

He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016. IEEE, 2016;770-778. DOI 10.1109/CVPR.2016.90

Huang M., Wang Q.G., Zhu Q.B., Qin J.W., Huang G. Review of seed quality and safety tests using optical sensing technologies. *Seed Sci. Technol*. 2015;43(3):337-366. DOI 10.15258/sst.2015.43.3.16

Jana B.K., Mukherjee S.K. Notes on the distribution of phytomelanin layer in higher plants – a short communication. *J. Pharm. Biol*. 2014;4(3):131-132

Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics. New York: Springer, 2002. DOI 10.1007/b98835

Khojastehnazhand M., Roostaei M. Classification of seven Iranian wheat varieties using texture features. *Expert Syst. Appl*. 2022;199: 117014. DOI 10.1016/j.eswa.2022.117014

Kolhar S., Jagtap J. Plant trait estimation and classification studies in plant phenotyping using machine vision. A review. *Inf. Process. Agric*. 2023;10(1):114-135. DOI 10.1016/j.inpa.2021.02.006

Komyshev E.G., Genaev M.A., Afonnikov D.A. Analysis of color and texture characteristics of cereals on digital images. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2020;24(4):340-347. DOI 10.18699/VJ20.626

Kozłowski M., Górecki P., Szczypiński P.M. Varietal classification of barley by convolutional neural networks. *Biosyst. Eng*. 2019;184: 155-165. DOI 10.1016/j.biosystemseng.2019.06.012

Li Z., Guo R., Li M., Chen Y., Li G. A review of computer vision technologies for plant phenotyping. *Comput. Electron. Agric*. 2020;176: 105672. DOI 10.1016/j.compag.2020.105672

Neuman M.R., Sapirstein H.D., Shwedyk E., Bushuk W. Wheat grain colour analysis by digital image processing II. Wheat class discri-

mination. *J. Cereal Sci*. 1989;10(3):183-188. DOI 10.1016/S0733-5210(89)80047-5

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. 2011;12:2825-2830

Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (Eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Vol. 9351. Cham: Springer, 2015;234-241. DOI 10.1007/978-3-319-24574-4_28

Sabanci K., Kayabasi A., Toktas A. Computer vision-based method for classification of wheat grains using artificial neural network. *J. Sci. Food Agric*. 2017;97(8):2588-2593. DOI 10.1002/jsfa.8080

Sharma R., Kumar M., Alam M.S. Image processing techniques to estimate weight and morphological parameters for selected wheat refractions. *Sci. Rep*. 2021;11(1):20953. DOI 10.1038/s41598-021-00081-4

Shoeva O.Yu., Strygina K.V., Khlestkina E.K. Genes determining the synthesis of flavonoid and melanin pigments in barley. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018;22(3):333-342. DOI 10.18699/VJ18.369 (in Russian)

Souza F.H., Marcos-Filho J. The seed coat as a modulator of seed-environment relationships in Fabaceae. *Braz. J. Bot*. 2001;24(4):365-375. DOI 10.1590/S0100-84042001000400002

Tan M., Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, 9–15 June 2019. ICML, 2019;6105-6114

van der Maaten L., Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res*. 2008;9(11):2579-2605.

Wang Y.H., Su W.H. Convolutional neural networks in computer vision for grain crop phenotyping: a review. *Agronomy*. 2022;12(11):2659. DOI 10.3390/agronomy12112659

Zykin P.A., Andreeva E.A., Tsvetkova N.V., Voylokov A.V. Anatomical and image analysis of grain coloration in rye. *Preprints*. 2020; 2020110530. DOI 10.20944/preprints202011.0530.v1

ORCID ID

D.A. Afonnikov orcid.org/0000-0001-9738-1409

Acknowledgements. The development of the phenotyping protocol, classification algorithm, and testing was financially supported by the Russian Science Foundation (project No. 22-74-00122, https://rscf.ru/project/22-74-00122/). For data analysis, computational resources of the Bioinformatics CPC were used with the support of budget project No. FWNR-2022-0020.

The authors would like to thank E.A. Zavarzin and A.I. Ivleva for their participation in training the neural network models.

Conflict of interest. The authors declare no conflict of interest.

Received June 30, 2023. Revised September 27, 2023. Accepted September 28, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):869-877
DOI 10.18699/VJGB-23-100

Original Russian text https://vavilovj-icg.ru/

Mathematical modeling of quorum sensing dynamics in batch culture of luminescent bacterium *Photobacterium phosphoreum* 1889

S.I. Bartsev[1, 2] ✉, A.B. Sarangova[2]

[1] Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
[2] Siberian Federal University, Krasnoyarsk, Russia
✉ bartsev@yandex.ru

Abstract. At the beginning of the paper, the level of necessary phenomenology of complex models is discussed. When working with complex systems, which of course include living organisms and ecological systems, it is necessary to use a phenomenological description. An illustration of the phenomenological approach is given, which captures the most significant general principles or patterns of interactions; the specific values of the parameters cannot be calculated from the first principles, but are determined empirically. An appropriate interpretation is also chosen empirically and pragmatically. However, in order to simulate a wider range of situations, it becomes necessary to lower the level of phenomenology, switch to a more detailed description of the system, introducing interaction between selected elements of the system. The requirements for a system model combining ecological, metabolic and genetic levels of cell culture description are formulated. A mathematical model of quorum sensing dynamics during the growth of batch culture of luminescent bacteria at different concentrations of the nutrient substrate has been developed. The model contains four blocks describing ecological, energy, quorum and luminescent aspects of bacterial culture growth. The model demonstrated good agreement with the experimental data obtained. When analyzing the model, three oddities in the behavior of the culture were noted, which presumably can change the idea of some processes taking place during the development of a culture of luminescent bacteria. The results obtained suggest the presence of some additional control system for the luminescent reaction via the synthesis pathways of $FMN \cdot H_2$ or aliphatic aldehyde. In this case, the generalized description of the contribution of energy metabolism to luminescence only through ATP is too strong a simplification. As a result of comparing the model dynamics with the experiment, a discrepancy arose between the concentration of the substrate (peptone) measured in the experiment and its effective influence on the bacterial population growth. This discrepancy seems to indicate peptone is not the leading substrate, and growth is limited by nutrients contained in the yeast extract, the concentration of which did not change in these experiments. The discrepancies noted between the expectations and the results of experimental data processing, together with the assumptions about the causes of these discrepancies, set the direction for further experimental and theoretical studies of quorum sensing mechanisms in a culture of luminescent bacteria
Key words: quorum sensing; mathematical model; luminescent bacteria.

For citation: Bartsev S.I., Sarangova A.B. Mathematical modeling of quorum sensing dynamics in batch culture of luminescent bacterium *Photobacterium phosphoreum* 1889. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):869-877. DOI 10.18699/VJGB-23-100

Математическое моделирование динамики кворум-эффекта в накопительной культуре люминесцентных бактерий *Photobacterium phosphoreum* 1889

С.И. Барцев[1, 2] ✉, А.Б. Сарангова[2]

[1] Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
[2] Сибирский федеральный университет, Красноярск, Россия
✉ bartsev@yandex.ru

Аннотация. В начале статьи обсуждается уровень необходимой феноменологичности сложных моделей. При работе со сложными системами, к которым, безусловно, относятся живые организмы и экологические системы, с необходимостью приходится использовать феноменологическое описание. Приведена иллюстрация феноменологического подхода, который ухватывает наиболее существенные даже не закономерности, а общие принципы или паттерны взаимодействий, причем конкретные значения параметров не могут быть вычислены из первых

принципов, а определяются эмпирически. Также эмпирически и прагматически выбирается соответствующая интерпретация. Однако для моделирования более широкого круга ситуаций возникает необходимость понижать уровень феноменологии, переходить на более детальное описание системы, вводя взаимодействие между выделенными элементами системы. Формулируются требования к модели системы, совмещающей экологический, метаболический и генетический уровни описания клеточной культуры. Разработана математическая модель динамики кворум-эффекта в процессе роста накопительной культуры люминесцентных бактерий при разных концентрациях питательного субстрата. Модель содержит четыре блока, описывающие экологический, энергетический, кворумный и люминесцентный аспекты развития культуры. Модель продемонстрировала хорошее соответствие экспериментальным данным, полученным в ходе выполнения работы. При анализе модели отмечены три странности в поведении культуры, которые, предположительно, могут изменить представление о некоторых процессах, имеющих место при развитии культуры люминесцентных бактерий. Полученные результаты позволяют предположить наличие некоторой дополнительной системы контроля люминесцентной реакции через пути синтеза ФМН·H_2 или алифатического альдегида. В этом случае обобщенное описание вклада энергетического метаболизма в люминесценцию только через АТФ является слишком сильным упрощением. В результате анализа результатов сопоставления модельной динамики с экспериментом возникло расхождение между измеряемой в эксперименте концентрацией субстрата (пептона) и его эффективным действием на рост популяции бактерий. Это расхождение, по-видимому, указывает на то, что пептон не является ведущим субстратом и рост лимитируют биогены, содержащиеся в дрожжевом экстракте, концентрация которого в этих экспериментах не изменялась. Отмеченные расхождения между ожиданиями и результатами обработки экспериментальных данных вместе с предположениями о причинах этих расхождений задают направление дальнейших экспериментальных и теоретических исследований механизмов кворум-эффекта в культуре люминесцентных бактерий.
Ключевые слова: кворум-эффект; математическая модель; люминесцентные бактерии.

Introduction

When working with complex systems, which of course include living organisms and ecological systems, it is necessary to use a phenomenological description. One of the widely used examples of a phenomenological description of a population is the Verhulst equation. Despite the fact that formally the population equations can be used only near the threshold of population survival (Gorban et al., 1982), this equation describes the dynamics of various processes quite well: the batch culture of microorganisms, the spread of an epidemic under constant conditions, population growth after invasion and the dynamics of sales under conditions of limited market capacity. Apparently this is due to the fact that at the final stage of the process, when the value of the variable approaches the carrying capacity, the specific growth rate approaches 0, which, in fact, corresponds to the approach to the threshold of survival.

Several versions of the Verhulst equation can be written, corresponding to different interpretations. Consider, for example, two of them:

$$\dot{N} = \mu_0(N_{\max} - N)N, \quad (1a)$$

$$\dot{N} = \mu_0 N - \alpha N^2. \quad (1b)$$

In the first version, $N_{\max}$ is called the carrying capacity, understood as the maximum population size that can exist under given conditions, and the product of $\mu_0 N_{\max}$ is the specific rate of population growth at a population size close to zero. The carrying capacity phenomenologically includes all kinds of factors limiting the growth of the population: substrate inhibition, inhibition by metabolites, limited plant growth area. This option corresponds well to the interpretation of plant or microbial population growth.

In the second version, μ_0 is the specific growth rate, α is a coefficient describing intraspecific competition, which can be realized by different mechanisms – competition for food and/or displacement from the hunting territory and direct collisions of individuals. This interpretation seems to be more appropriate for animals.

These examples are given to illustrate the phenomenological approach, which captures the most significant, not even regularities, but general principles or patterns of interactions, and the specific values of the parameters cannot be calculated from the first principles, but are determined empirically. The appropriate interpretation is also chosen empirically and pragmatically.

However, to model a wider range of situations, there is a need to lower the level of phenomenology, move to a more detailed description of the system, introducing interaction between selected elements of a system. For example, there are cases when the Verhulst equation does not describe the dynamics of a batch culture accurately enough. In this case, it is necessary, for example, to take into account substrate dynamics and introduce substrate inhibition of culture growth. At the same time, we still remain at a very high level of phenomenology, continuing to describe the dependence of culture growth using the Monod formula and its various modifications and complications, reducing the entire metabolism of a cell or multicellular organism to one key enzymatic reaction.

The need to lower the level of phenomenology arises when the researcher encounters phenomena that do not fit into the existing model. In this case, it is often necessary to move to the level of genetic and/or metabolic regulation of cellular processes. One such example that requires reducing the phenomenological nature of the models used is quorum sensing (QS) (Miller, Bassler, 2001). It is noteworthy and symbolic that QS, which is a manifestation of molecular-level events at the population level, was discovered in luminescent bacteria, the luminescence of which is a natural indicator of the current state of cellular metabolism (Nealson et al., 1970).

The quorum sensing is the expression of certain genes being triggered when a certain threshold population density is reached. At the bacterial level, this effect is based on the synthesis and release into the external environment of signal molecules (autoinducers), the concentration of which varies depending on the number of surrounding cells, and, when a

certain threshold concentration is exceeded, the expression of certain genes is triggered. Since QS occurs in a fairly wide range of organisms (for example, insects (Anstey et al., 2009) and fish (Makris et al., 2009)), its study seems quite important in itself. In addition, identifying the patterns of manifestation of QS and its prediction is important for the microbiological synthesis of products triggered by this effect. An example of such a product is bacterial luciferase, which is used for laboratory and rapid toxicological biotests. At the same time, luminescent bacteria are a convenient tool for studying QS, since luminescence is a natural function of cells, which makes it possible to study the process on native cells without the introduction of special fluorescent dyes and without stimulating fluorescence. The evolutionary meaning of QS in luminescent bacteria is explained within the framework of the hypothesis that the selection mechanism is associated with spread and reproduction of bacteria (Nealson, Hastings, 1979). As marine enterobacteria, luminescent bacteria growing on a substrate (the surface of dead organisms or fecal pellets), if the culture density is sufficient, can produce enough light to attract organisms to consume them, thereby ensuring the circulation of bacteria through the intestinal tracts of sea animals.

The purpose of this work is to develop a mathematical model and its software implementation for the analysis of experimental data on QS in batch culture of luminescent bacteria. To specify the requirements for the model, we will formulate a kind of technical specification (TS) for the model being developed. Firstly, the model must describe the dynamics of bacterial growth in batch culture; secondly, it must describe the dynamics of the luminescence of a bacterial culture, which is regulated by QS, i. e. events at the molecular level; thirdly, the model should be as simple as possible for the simple reason that a complex model contains a large number of parameters with unknown values, i. e. we follow the paradigm that the fewer fitting parameters there are in a model describing complex processes, the more it reflects the essence of the processes being modeled.

The third point of our conditional TS mentions the complexity of the model, and since this point demands the simplicity of the model being created, at least a brief discussion of this term is required. Unfortunately, there is no universal definition of complexity; this is evidenced by the huge (>40) number of existing definitions of complexity (Edmonds, 1999). The peculiarities of applying this term to the description of evolving living systems make it possible to narrow down the set of possible definitions (Bartsev, Bartseva, 2010). In the case of mathematical models constructed as systems of ordinary differential equations (ODEs), often used to describe the chemical (biochemical) kinetics and dynamics of ecological systems, a natural (or at least widely used) indicator of complexity is the number of differential equations in the system. Apparently, it is not for nothing that methods that make it possible to reduce the dimension of an ODE system, for example, by selecting a subsystem of fast motions and applying Tikhonov's theorem (Romanovsky et al., 1984), are called methods for simplifying systems of kinetic equations.

True, the question remains about the complexity of the equations themselves, or rather, their right-hand sides. It is obvious that functions including a larger number (so to speak) of nonlinearities, for example, terms with large powers in a fractional rational function, can provide more diverse behavior. A possible quantitative approach to assessing the complexity of ODE systems, taking into account the degree of nonlinearity of the right-hand sides, can be based on Korzukhin's theorem (Jabotinsky, 1974). It states that for a system with nonlinear right-hand sides, a system of chemical kinetics equations (containing terms that describe reactions no higher than second order) can be constructed so that the behavior of some of the variables of the new system will coincide with the behavior of the variables of the original one. The number of equations of the second, expanded system could serve as a measure of the complexity of the model, taking into account the degree of nonlinearity of the right-hand sides used. Since our task is not to obtain an accurate estimate of the model complexity, but only to construct the simplest possible model that provides an adequate description of the real system, we will simply minimize the number of differential equations of the model and simultaneously use the minimum degrees of variables in there right-hand sides.

Methods and materials

Experimental part. The object of the study are luminous bacteria *Photobacterium phosphoreum* 1889, from the collection of the Institute of Biophysics SB RAS. Bacterial growth was assessed by measuring optical density at 660 nm on an Agilent Cary 60 spectrophotometer. To measure the bioluminescence of the reaction mixture Promega GloMax 20/20 Luminometer (USA) was used. The bacteria were grown in a liquid medium for marine bacteria (g/l): NaCl – 28.5, KCl – 0.5, $CaCl_2$ – 0.5, $MgCl_2$ – 4.5, yeast extract – 1, peptone – 10; pH 7.6.

Mathematical model. The bioluminescent system of bacteria has been very well studied (Brodl et al., 2018), the enzymes expressed jointly when QS is triggered are known, and the pathways for the synthesis of substrates for the luminescent reaction are quite well studied. For us, in order not to dive into the details of the kinetics of the multienzyme system, the following is important: the direct substrates of the luminescent reaction are reduced flavin mononucleotide ($FMN \cdot H_2$), long-chain aliphatic aldehyde – tetradecanal and molecular oxygen. The flavin is reduced by the enzyme NADH:FMN oxidoreductase, and the aldehyde is synthesized by the ATP-consuming fatty acid reductase enzyme complex. Thus, the luminescent reaction is directly related to the energy metabolism of the cell, and its luminescence depends not only on the amount of luciferase in the cell, but also on the state of its energy metabolism.

Consequently, already at the level of describing crop growth, an assessment of the state of its energy metabolism must be included in the model. The properties of the multienzyme system of energy metabolism were studied in detail in the almost forgotten (judging by the citation statistics from ResearchGate) work of E.E. Selkov, which is part of a collective monograph (Ivanitsky et al., 1978). One of the most important properties of energy metabolism is maintaining a constant intracellular ATP concentration within a wide range of consumer load to ensure decoupling (relative independence) of intracellular energy consumers. In his work, the case of a constant rate of

substrate supply with varying load (activity of generalized ATPase) was considered. In this model, it is necessary to take into account both the change in the rate of substrate supply (in our case, we will consider its concentration in the medium) and changes in ATPase activity associated with different phases of culture growth. Here the entire Selkov's model, in accordance with TS-3, will not be reproduced, but some of his ideas will be applied.

When writing a model that, on the one hand, describes the variables characterizing a bacterial culture – substrate concentration and biomass density in the flask, and, on the other hand, should describe the average intracellular ATP concentration, it is necessary to coordinate the rates of the processes. If we designate the volume of the flask as V_c, and the total volume of bacterial cells as V_b, then between the rates of processes expressed in concentrations per unit of time – v_c and v_b, respectively, due to the conservation law, the following relation must be satisfied: $v_c \cdot V_c = v_b \cdot V_b$, where the right and left sides of the equality describe the rate of change in the mass of the reagent. It follows that the rates of intracellular processes must exceed the (concentration) rates of the same processes by V_c/V_b times and we will have a system with different characteristic times of change in variables. Let us denote the ratio V_b/V_c as a small parameter ε_0.

Taking into account the above, the "ecological part" of the model can be written as follows:

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = [f_G(S,a) - M_N(a)]N, \\ \varepsilon_0 \dot{a} = f_E(S,a) \cdot \dfrac{N}{\varepsilon_1 + N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2 + a}, \end{cases} \quad (2)$$

where S is the concentration of the nutrient substrate; N is the bacterial biomass; a is the average intracellular concentration of ATP in the cells of a bacterial culture.

In this case, the function $f_G(S,a) = \dfrac{V_G S}{K_G + S} \cdot \dfrac{a}{K_{aG} + a}$ describes the ATP-dependent synthesis of biomass, the function $f_E(S,a) = \dfrac{V_E S}{K_E + S} \cdot \dfrac{a}{K_{aE} + a^2}$ describes the production of ATP, the expression $\dfrac{k_d a}{\varepsilon_2 + a}$ describes the activity of the generalized ATPase, and the function $M_N(a) = \dfrac{m}{1 + A_N a}$ describes the intensity of bacterial death, depending on the intracellular ATP concentration.

As one can see, in this model the generalized activities of anabolic and catabolic pathways are described by separate functions, so there is no need to specifically introduce the so-called economic coefficient; moreover, the ratio of the rates of biomass synthesis and organic oxidation may change during crop growth. The type of function $f_E(S, a)$, or more precisely its part, describing the dependence of the activity of ATP synthesis on its concentration, was chosen in accordance with Selkov's model (Ivanitsky et al., 1978). The last term in the equation describing the ATP concentration represents the contribution of the generalized ATPase, i. e. the totality of all basic processes in a cell. At small values of the coefficient ε_2 ATPase activity will change little over a wide range of ATP concentrations, and only at low values a drop in ATPase activity will be observed, which seems natural.

The presence of a small parameter in the third equation makes the ATP concentration a fast variable and allows us to study the properties of this equation separately from other variables, assuming the remaining (ecological) variables are constants (Romanovsky et al., 1984). We will not do a complete analysis of the stability of this equation due to its cumbersomeness; it is enough for us, in accordance with the technical specifications, to check the possibility of the existence of a stable quasi-stationary state of a given dynamic system and evaluate the dependence of its stability on the values of environmental variables.

From Figure 1 we can see that depending on the set of parameters the system can have: (*A*) one stable zero stationary state, or three stationary states depending on the substrate concentration *S*; (*B*) one unstable zero and one stable stationary state for any values of concentration *S*. Since at this stage we are not concerned with the exact correspondence of the parameters of the cell energy system model to real data, we will follow Selkov's approach and the stated technical specifications. It means we will choose an option, on the one hand, providing the cell with stable satisfaction of its energy needs, and on the other, doing this in the simplest way. From Figure 1 it is clear that this requirement is met by a set of parameters that generates the dependencies presented in the sub-figure (*B*).

It can be seen from the figure that at certain parameter values there is a range of changes in the ATP concentration, in which the rate of ATP synthesis is positive, which leads to an increase in its concentration until the concentration falls into the region with a negative rate value, which ensures the existence of a stable stationary state.

Having ensured, relatively speaking, the vital activity of the cell, we can move to constructing a model of QS. Let us consider the QS model (Williams et al., 2008), which was subsequently used in a number of works by other authors (Melke et al., 2010; Djezzar et al., 2019). According to this model, the autoinducer *AHL* (*A*) and the receptor *LuxR* (*R*) form a dimerized complex that regulates the production of both *R* and *A*. In addition, there is a nonzero, basal, inducer concentration-independent synthesis of *LuxR*. The model looks like this:

$$\begin{cases} \dot{R} = C_R + \dfrac{V_R D}{K_R + D} - k_3 R - k_1 R A + k_2 C, \\ \dot{C} = k_1 R A - k_2 C - 2k_4 C^2 + 2k_5 D, \\ \dot{D} = k_4 C^2 - k_5 D. \end{cases} \quad (3)$$

In this system, the first equation describes the rate of change in the *LuxR* concentration, which positively depends on the sum of the basal (C_R) and autoinduced synthesis rates. The latter is proportional to the probability of transcription initiation controlled by binding the $(LuxR\text{-}A)_2$ (*D*) complex to the corresponding binding site in the regulatory sequence of the operon. The second and third equations describe the formation of the *LuxR-A* complex (*C*) followed by the formation of the dimeric complex $(LuxR\text{-}A)_2$ (*D*).

Following (Williams et al., 2008) and TS-3, we will assume the existence of a quasi-stationary state for variables *C*

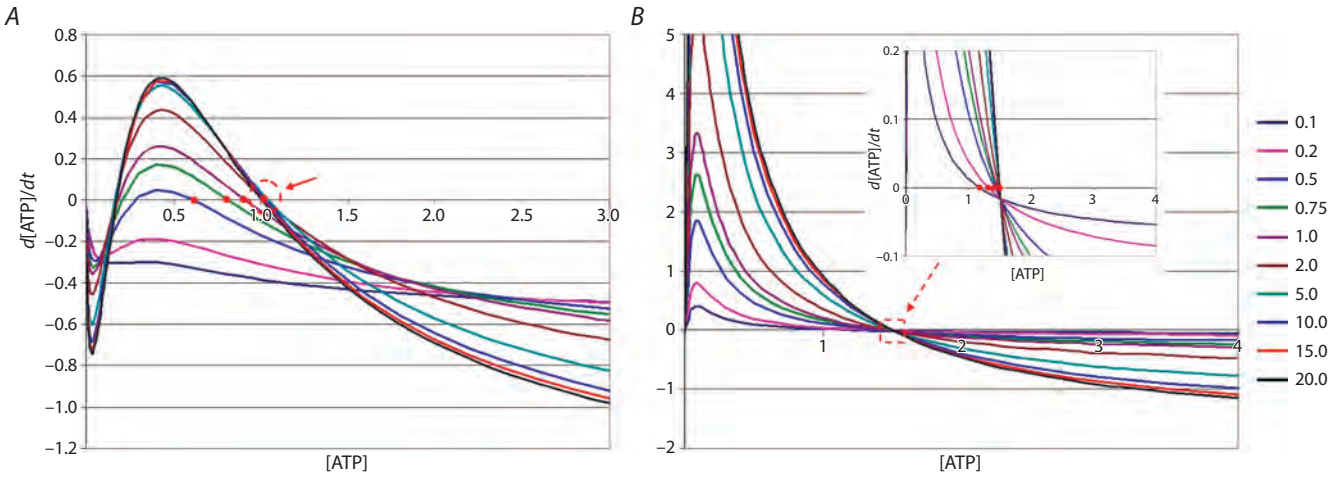


Fig. 1. Dependence of the rate of ATP concentration on its concentration at different substrate concentrations (shown on the right) at different sets of parameter values.

Case *A*: in a system at $S > S_{min}$, there may be three stationary states, of which one is unstable, and one corresponds to a zero concentration of ATP. The dashed oval highlights a group of indistinguishable stationary states at different values of *S*. Case *B*: there is one stable and one unstable zero stationary state in the system. The red circles show stable stationary states at different substrate concentrations. Appropriate parameter sets for cases *A*: $V_g = 1.22$, $K_g = 1.94$, $K_a = 0.01$, $V_e = 2$, $K_e = 1$, $K_{ae} = 0.2$, $k_d = 0.5$, $\varepsilon_2 = 0.05$ and *B*: $V_g = 2.18$, $K_g = 4$, $K_a = 0.004$, $V_e = 3.299$, $K_e = 4$, $K_{ae} = 0.008$, $k_d = 0.026$, $\varepsilon_2 = 0.85$.

and *D*. Then, the equation describing the behavior of *LuxR* at the concentration of the autoinducer considered as an external parameter has the form:

$$\dot{R} = C_R + \frac{V_R \gamma R^2 A^2}{K_R + \gamma R^2 A^2} - k_3 R, \qquad (4)$$

where $\gamma = \dfrac{k_4 k_1^2}{k_5 k_2^2}$.

To analyze the properties of this equation, one can apply the technique used for the third equation of system (2), that is, consider it in coordinates (*R*, *dR*/*dt*) at different concentrations of the autoinducer, which is a simple matter (Fig. 2). It can be seen from the figure that at zero and low substrate concentrations, only one stable stationary state can exist, corresponding to a low *LuxR* concentration. As the concentration of the autoinducer increases, two more stationary states appear – stable and unstable, but the system cannot voluntarily switch to a state with a high level of *LuxR* expression. With a further increase in the concentration of the autoindicator, the left knee of the curve leaves the negative half-plane, which leads to the disappearance of the unstable and stable states and the system quickly transitions to a state with a high concentration of *LuxR*.

The switching process can be shown more clearly if we assume that a quasi-stationary state of the system described by equation (4) is realized. In this case, one can either apply plotting implicitly defined functions in computer algebra systems like Maxima, or, by setting the right side equal to 0, obtain an expression for the explicit function

$$A = \frac{1}{R}\sqrt{\frac{\sigma(R-\alpha)}{(\alpha+\beta)-R}}, \qquad (5)$$

where $\sigma = K_R\gamma$; $\alpha = C_R/k_3$; $\beta = V_R/k_3$. In this case, the condition $\alpha < R < \alpha + \beta$ must be satisfied.

For clarity, one can tabulate (5) as a regular function in Excel, and then flip the coordinates – make (*A*, *R*) (Fig. 3).

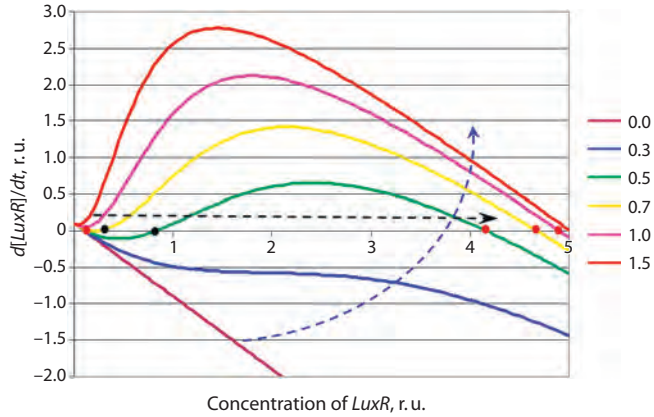


Fig. 2. Dependence of the rate of change in the concentration of *LuxR* on its concentration at different concentrations of the auto-inductor shown on the right.

The red circles show stable stationary states at different concentrations of the auto-inductor, the black ones show unstable states. The curved dashed arrow indicates the direction of change in the concentration of the auto-inductor. The straight dashed arrow shows the direction of switching to the new state.

The figure clearly shows that when a certain threshold concentration of $A(\beta)$ is exceeded, a sharp transition to a state of high level of *LuxR* expression occurs, and hysteresis can be observed in the system, which under natural conditions can be observed when bacterial growth is inhibited and the autoinducer is gradually destroyed.

After running the QS model and tentatively estimating the values of the parameters that are necessary to implement QS, we will return to building the total model. Using the well-known model discussed above, we will slightly modify it to ensure its conceptual unity, namely, we will make the intensive synthesis of *LuxR* energy-dependent. In this case, we will leave the background synthesis of the autoinducer

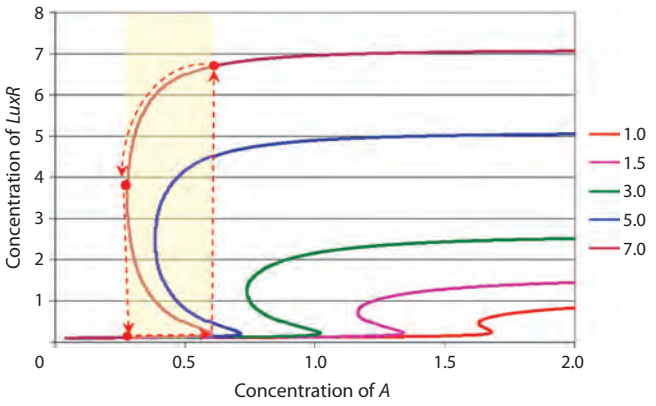


Fig. 3. Stationary curves showing the dependence of the stationary *LuxR* concentration on the concentration of the auto-inductor at α = 0.1, σ = 1 and different values of the parameter β (right).

and *LuxR* conditionally energy-independent, considering that the costs of their synthesis are included in the activity of the generalized ATPase (2):

$$\begin{cases} \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R \gamma R^2 A^2}{K_R + \gamma R^2 A^2} \cdot \dfrac{a}{\varepsilon_3 + a} - k_3 R. \end{cases} \quad (6)$$

Looking ahead, we can say that the use of a more complex equation, assuming that simultaneously with the synthesis of *LuxR*, the synthesis of the autoinducer is intensified, as was done in the model (Melke et al., 2010), turned out to be unnecessary to describe the experimental data. In addition, for simplicity, it is assumed that the concentration of the autoinducer in the medium and in the cell coincide, which makes it possible to avoid selecting a small parameter. As a result, our model, combining environmental and intracellular molecular processes, looks like this:

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = [f_G(S,a) - M_N(a)]N, \\ \varepsilon_0 \dot{a} = f_E(S,a) \cdot \dfrac{N}{\varepsilon_1 + N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2 + a}, \\ \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R \gamma R^2 A^2}{K_R + \gamma R^2 A^2} \cdot \dfrac{a}{\varepsilon_3 + a} - k_3 R. \end{cases} \quad (7)$$

Let's start constructing the final luminescent block of the model. First, let us assume that the synthesis of luciferase occurs in parallel with the synthesis of *LuxR* and is also energy dependent. In addition, we will take into account the energy-independent process of luciferase inactivation. However, in this experiment we record not the amount of luciferase in the culture, but the intensity of luminescence. As mentioned above, to ensure luminescence, NADH and ATP must come from the cell. It is possible to take these flows into account separately, but it hardly makes sense, since the activity of the cytochrome chain that produces ATP depends on the presence of NADH. Thus, since these processes are closely related and the presence of ATP means the presence of NADH, in the model (following TS-3) we will consider the dependence of luminescence only on ATP. As a result, we obtain a general model of the system under consideration:

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = f_G(S,a)N, \\ \varepsilon_0 \dot{a} = f_E(S,a) \cdot \dfrac{N}{\varepsilon_1 + N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2 + a}, \\ \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R \gamma R^2 A^2}{K_R + \gamma R^2 A^2} \cdot \dfrac{a}{\varepsilon_3 + a} - k_3 R, \\ \dot{L} = \dfrac{V_L R}{K_L + R} \cdot \dfrac{a}{\varepsilon_3 + a} - k_{dL} L, \end{cases} \quad (8)$$

$$Light(t) = L(t) \cdot \frac{a(t)}{\varepsilon_4 + a(t)}.$$

The difference between the ecological part of this model and (2) is that since the experiment considers batch culture from inoculation to the logarithmic growth phase inclusively, without considering the stationary phase and the death phase, the mortality of bacteria cannot be taken into account in this experiment.

The mathematical model was implemented in the open source environment SciLab 6.1. To determine the parameters of the mathematical model from experimental data, the Nelder–Mead method was used, the code of which is included in the accompanying software examples.

Results

Test experiments carried out on the prescribed rich (10 g/l peptone) and poor media (0.1 g/l peptone) showed the presence of QS in both cases. The biomass and luminescence dynamics curves are shown in Figure 4.

Simply examining the obtained curves, without any model, one can see (see Fig. 4, *a*) that before the start of QS (within 6 hours of cultivation) there is a gradual decrease in the intensity of luminescence produced by luciferase brought with the inoculum. At the same time, after reaching the maximum of the glow (~11 hours), a sharp decrease in the intensity of the glow is observed. It is almost obvious that such a decline cannot be associated with inactivation of luciferase, which would require the assumption of the existence of a special system that destroys luciferase immediately after synthesis, and even under conditions of energy starvation. Apparently, it was the drop in the concentrations of NADH and ATP at the final stage of the logarithmic phase of culture growth that caused this drop in luminescence. At the same time, the slow decrease in luminescence intensity, which took place under conditions of excess substrate and intensive energy metabolism, demonstrates the process of inactivation of luciferase, or more precisely the complex of enzymes that serve the luminescence of bacteria.

At the same time, the dynamics of cultural luminescence in a poor environment (see Fig. 4, *b*) raises interesting questions. It can be seen that after 7 hours of cultivation, the biomass of bacteria reached approximately more than a third of the biomass achieved by bacteria during the same time in the rich medium. At the same time, the growth rate of the culture, although not very high, was approximately constant throughout the entire period under consideration, which cannot be

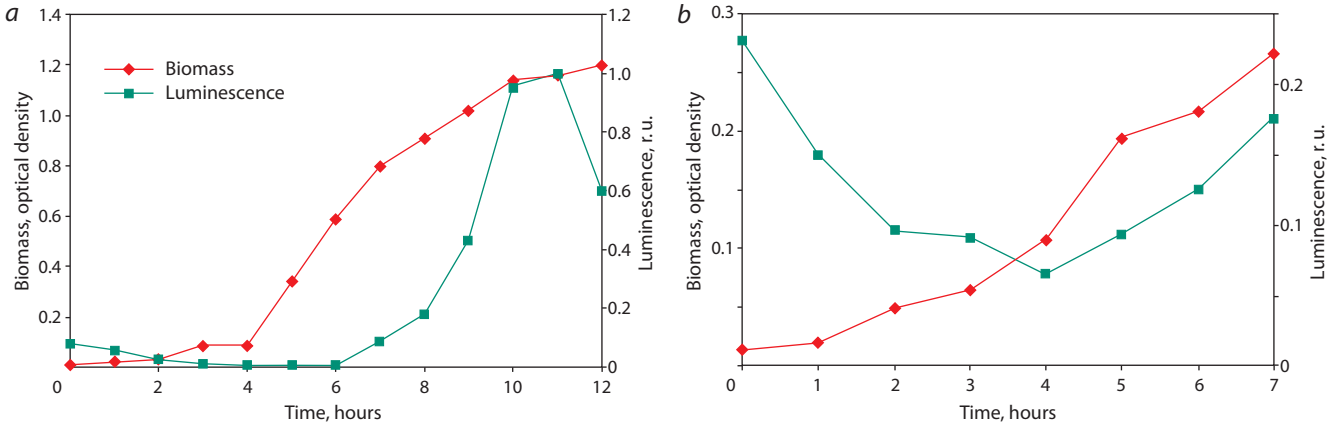


Fig. 4. Dynamics of biomass growth and luminescence of *Photobacterium phosphoreum* 1889 culture on a standard medium (*a*) and on a poor medium (*b*).

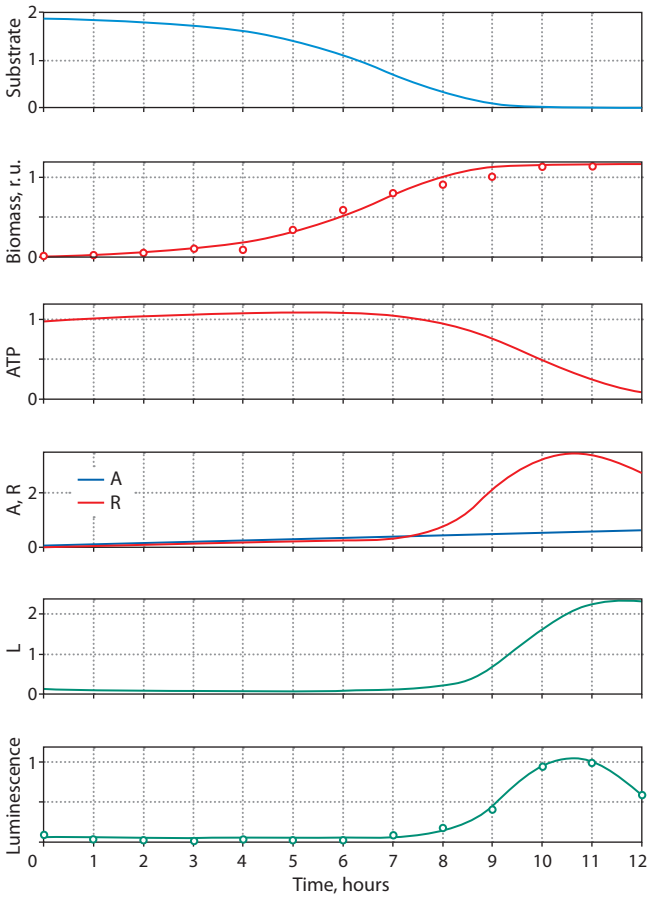


Fig. 5. Model and real dynamics of variables in the culture of luminescent bacteria.

Circles indicate experimental data.

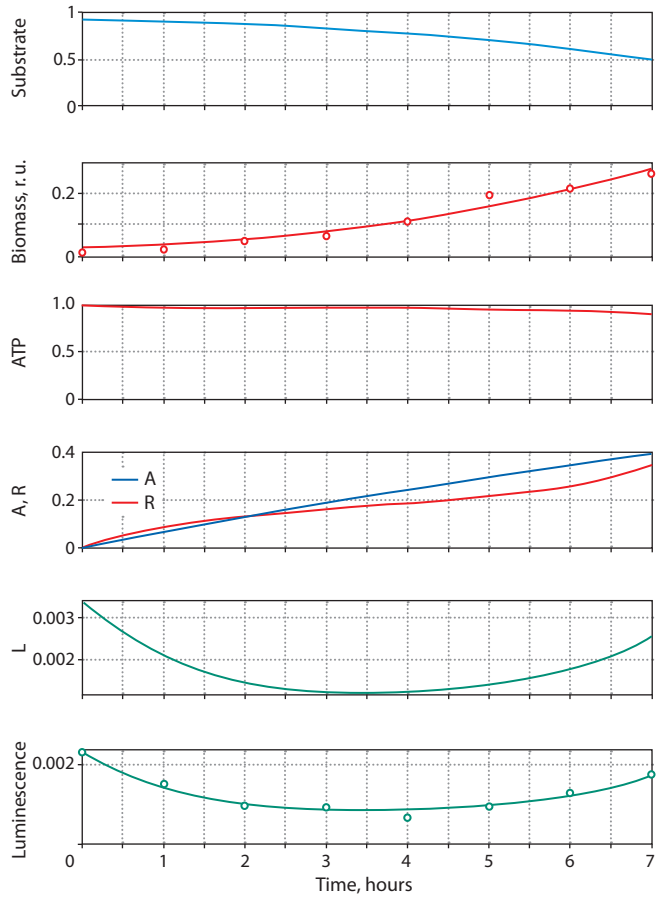


Fig. 6. Model and real dynamics of variables in the culture of fluorescent bacteria in a poor environment.

Circles indicate experimental data.

said about the other experiment. The obvious acceleration of culture growth in a rich medium after 4 hours of growth may indicate substrate inhibition at given substrate concentrations.

It is interesting that in the poor medium QS began 2 hours earlier than in the rich medium. It is possible that substrate inhibition has this effect, but this issue requires further research and more experimental material. At this stage, our objective is to develop an adequate model that satisfies the technical specifications stated at the beginning of the article, and to preliminarily test the adequacy of this model using the available experimental data.

The results of computational modeling are shown in Figures 5 and 6. The adjustment of the model parameters took place in two stages – first, the ecological part was adjusted, describing the dynamics of the biomass of the bacterial culture, substrate concentration and the average intracellular ATP

concentration. The results are shown in the top three graphs of the presented figures. It should be noted that there is fairly good agreement between the model curve describing the dynamics of biomass and the experimental points. As additional calculations have shown, neither the Verhulst model nor the introduction of the substrate inhibition factor into the model provides an improved description. Two options are possible: either the observed discrepancy is of a statistical nature, or there is a mechanism in the system that accelerates growth after reaching a certain threshold. Further analysis will require additional experiments, which are planned.

It should be noted the expected behavior of the ATP concentration, which, as can be seen from Figure 1, should undergo minor changes when the substrate concentration varies in a certain interval and change quite sharply when leaving this interval.

At the second stage, the part of the model describing QS and luminescence was adjusted, and data on luminescence dynamics were used as reference data. At the same time, the "ecological and energy" parameters of the model did not change.

Figure 5 shows the model dynamics of the autoinducer and *LuxR*, as well as the dynamics of the amount of luciferase, which follows the dynamics of *LuxR* expression. It is important to note that the model clearly describes the slow decrease in luminescence at the initial stage of culture growth and its rapid decrease at the final stage, which differs in rate from the decrease in the amount of luciferase, which represents the energy state of the cells.

In the case of modeling the behavior of a culture in a poor environment (see Fig. 6), the following point can be noted. It is important that a model containing a large number of adjustable parameters is capable of describing various variants of dynamics, and the question is how well these parameters correspond to biological ideas about the system under study. Looking at the graphs in Figure 6, one can note good agreement between the model curves and the experimental data. Let us compare in the "Discussion" section the changes in the constants that were made by the system for adjusting parameters when describing the growth of a culture in a poor environment. In this case, the values of the model parameters common for the two cases are as follows: $V_g = 2.18$, $K_g = 3.99$, $K_a = 0.0033$, $V_e = 3.30$, $K_e = 4.02$, $K_{ae} = 0.008$, $a_0 = 1.40$, $k_d = 0.0315$, $k_0 = 0.082$, $V_R = 1.50$, $C_A = 0.14$, $C_R = 0.011$, $k_3 = 0.057$, $\gamma = 0.331$, $K_R = 0.06$, $K_L = 0.17$, $\varepsilon_0 = 0.01$, $\varepsilon_1 = 0.001$, $\varepsilon_2 = 1.54$, $\varepsilon_3 = 0.39$, $\varepsilon_4 = 3.34$.

Discussion

From the Table it can be seen that it was not necessary to change a very large number of parameters in order to obtain a good description of the dynamics of culture in both experiments. Note that the change in S_0 is expected; another thing is that the almost twofold decrease in S_0 in the model is in poor agreement with a hundredfold decrease in the peptone concentration in the medium. This discrepancy can be tentatively explained by the fact that, apparently, peptone is not the leading substrate and growth is limited by nutrients contained in the yeast extract, the concentration of which did not change in these experiments.

Comparison of model parameters for two types of nutrient medium

Medium	S_0	k_{dL}	V_L
Rich	1.84	0.75	9.02
Poor	1.06	0.41	0.15

Questions arise regarding changes in the other two parameters. Such a significant (60 times!) drop in the V_L constant can only be explained by the presence of some additional system for controlling the luminescent reaction through the synthesis pathways of $FMN \cdot H_2$ or aliphatic aldehyde. In this case, a generalized description of the contribution of energy metabolism only through ATP is too strong a simplification.

The almost twofold decrease in the k_{dL} constant during growth in a poor medium is also difficult to explain. It is premature to build hypotheses on this matter; we can return to the issue after obtaining additional experimental data.

The noted discrepancies between expectations and the results of processing experimental data, together with assumptions about the courses of these discrepancies, set the direction for further experimental and theoretical studies of the mechanisms of QS in the culture of luminescent bacteria.

Conclusion

The results of a comparison of the model built within the framework of the presented logic and experimental data show that the proposed model generally satisfies the conditional technical specifications that were formulated in the Introduction. Indeed, (1) the model quite satisfactorily describes the dynamics of bacterial biomass in batch culture, (2) the model clearly describes the dynamics of the luminescence of a bacterial culture, which is regulated by QS.

But regarding the third requirement of the technical specifications about the maximum simplicity of the model, it is difficult to give a final assessment. On the one hand, it is possible that this model can be simplified to describe the behavior of bacterial cultures under conditions close to the conditions of the experiments considered. On the other hand, working with the model (selection of parameters) made us feel that this model is not robust enough with respect to the variation of parameters. This was manifested, in particular, in the fact that the Nelder–Mead method, like any local search method, quite often finds the nearest minimum of the goal function, which corresponds to the values of parameters that are distantly related to biological meaning (the tendency of the Monod constant to 0). It is possible that a model, in which the semantic blocks (ecological, energy, quorum, luminescent) will be more articulated, more autonomous, in line with the ideas of E.E. Selkov, will be resistant to external and internal disturbances, almost like a living being.

References

Anstey M.L., Rogers S.M., Ott S.R., Burrows M., Simpson S.J. Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *Science*. 2009;323(5914):627-630. DOI 10.1126/science.1165939

Bartsev S.I., Bartseva O.D. Heuristic Neural Network Models in Biophysics: Application to the problem of structure–function mapping. Krasnoyarsk: Siberian Federal University Publ., 2010 (in Russian)

Brodl E., Winkler A., Macheroux P. Molecular mechanisms of bacterial bioluminescence. *Comput. Struct. Biotechnol. J.* 2018;16:551-564. DOI 10.1016/j.csbj.2018.11.003

Djezzar N., Pérez I.F., Djedi N., Duthen Y. A computational multiagent model of bioluminescent bacteria for the emergence of self-sustainable and self-maintaining artificial wireless networks. *Informatica.* 2019;43(3):395-408. DOI 10.31449/inf.v43i3.2381

Edmonds B. Syntactic Measures of Complexity. Doctoral Thesis. Manchester, UK: Univ. of Manchester, 1999.

Gorban A.N., Okhonin V.A., Sadovskiy M.G., Khlebopros R.G. The simplest equation of mathematical ecology. Preprint of the Sukachev Forest and Timber Institute, Siberian Branch of the USSR Academy of Sciences, 1982 (in Russian)

Ivanitsky G.R., Krinsky V.I., Selkov E.E. Mathematical Biophysics of the Cell. Moscow: Nauka Publ., 1978 (in Russian)

Jabotinsky A.M. Concentration Oscillations. Moscow: Nauka Publ., 1974 (in Russian)

Makris N.C., Ratilal P., Jagannathan S., Gong Z., Andrews M., Bertsatos I., Godø O.R., Nero R.W., Jech J.M. Critical population density triggers rapid formation of vast oceanic fish shoals. *Science*. 2009;323(5922):1734-1737. DOI 10.1126/science.1169441

Melke P., Sahlin P., Levchenko A., Jönsson H. A cell-based model for quorum sensing in heterogeneous bacterial colonies. *PLoS Comput. Biol*. 2010;6(6):e1000819. DOI 10.1371/journal.pcbi.1000819

Miller M.B., Bassler B.L. Quorum sensing in bacteria. *Annu. Rev. Microbiol.* 2001;55(1):165-199. DOI 10.1146/annurev.micro.55.1.165

Nealson K.H., Hastings J.W. Bacterial bioluminescence: its control and ecological significance. *Microbiol. Rev.* 1979;43(4):496-518. DOI 10.1128/mr.43.4.496-518.1979

Nealson K.H., Platt T., Hastings J.W. Cellular control of the synthesis and activity of the bacterial luminescent system. *J. Bacteriol.* 1970; 104(1):313-322. DOI 10.1128/jb.104.1.313-322.1970

Romanovsky Yu.M., Stepanova N.V., Chernavsky D.S. Mathematical Biophysics. Moscow: Nauka Publ., 1984 (in Russian)

Williams J.W., Cui X., Levchenko A., Stevens A.M. Robust and sensitive control of a quorum-sensing circuit by two interlocked feedback loops. *Mol. Syst. Biol.* 2008;4:234. DOI 10.1038/msb.2008.70

ORCID ID

S.I. Bartsev orcid.org/0000-0003-0140-4894

Acknowledgements. The study was funded by State Assignment of the Ministry of Science and Higher Education of the Russian Federation (project No. 0287-2021-0018).

Conflict of interest. The authors declare no conflict of interest.

Received July 18, 2023. Revised September 16, 2023. Accepted September 16, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):878-883
DOI 10.18699/VJGB-23-101

Original Russian text https://vavilovj-icg.ru

Alga-based mathematical model of a life support system closed in oxygen and carbon dioxide

D.A. Semyonov ✉, A.G. Degermendzhi

Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
✉ semenov@ibp.ru

Abstract. The purpose of the study was to compare quantitative analysis methods used in the early stages of closed-loop system prototyping with modern data analysis approaches. As an example, a mathematical model of the stable coexistence of two microalgae in a mixed flow culture, proposed by Bolsunovsky and Degermendzhi in 1982, is considered. The model is built on the basis of a detailed theoretical description of the interaction between species and substrate (in this case, illumination). The ability to control the species ratio allows you to adjust the assimilation quotient (AQ), that is, the ratio of carbon dioxide absorbed to oxygen released. The problem of controlling the assimilation coefficient of a life support system is still relevant; in modern works, microalgae are considered as promising oxygen generators. At the same time, modern works place emphasis on empirical modeling methods, in particular, on the analysis of big data, and the work does not go beyond the task of managing a monoculture of microalgae. In our work, we pay attention to three results that, in our opinion, successfully complement modern methods. Firstly, the model allows the use of results from experiments with monocultures. Secondly, the model predicts the transformation of data into a form convenient for further analysis, including for calculating AQ. Thirdly, the model allows us to guarantee the stability of the resulting approximation and further refine the solution by small corrections using empirical methods.
Key words: life support system (LSS); mathematical model; mixed culture of two algae.

For citation: Semyonov D.A., Degermendzhi A.G. Alga-based mathematical model of a life support system closed in oxygen and carbon dioxide. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7): 878-883. DOI 10.18699/VJGB-23-101

Математическая модель системы жизнеобеспечения на основе водорослей, замкнутая по кислороду и углекислому газу

Д.А. Семёнов ✉, А.Г. Дегерменджи

Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
✉ semenov@ibp.ru

Аннотация. Целью исследования было сравнить методы количественного анализа, применявшиеся на ранних этапах создания прототипов замкнутых систем, с современными подходами анализа данных. В качестве примера рассмотрена математическая модель устойчивого сосуществования двух микроводорослей в смешанной проточной культуре, предложенная Болсуновским и Дегерменджи в 1982 г. Модель построена на основе детального теоретического описания взаимодействия видов и субстрата (в данном случае освещенности). Возможность управления соотношением видов позволяет регулировать ассимиляционный коэффициент (AQ), т. е. отношение поглощенного углекислого газа к выделенному кислороду. Задача управления ассимиляционным коэффициентом системы жизнеобеспечения до сих пор актуальна, микроводоросли рассматриваются как перспективные генераторы кислорода и в современных работах. При этом акцент в них сделан на эмпирических методах моделирования, в частности на анализе больших данных; также работы не выходят за пределы задачи управления монокультурой микроводорослей. В настоящем исследовании мы обращаем внимание на три результата, по нашему мнению, удачно дополняющих современные методы. Во-первых, модель позволяет использовать результаты экспериментов с монокультурами, во-вторых, предсказывает преобразование данных к виду, удобному для дальнейшего анализа, в том числе для вычисления AQ. В-третьих, модель позволяет гарантировать устойчивость полученного приближения и в дальнейшем искать решение как малую поправку эмпирическими методами.
Ключевые слова: система жизнеобеспечения (СЖО); математическая модель; смешанная культура двух водорослей.

Introduction

Nowadays, complex systems are predominantly viewed as a "black box" generating large amounts of data. The development of relevant methods for big data analysis has been facilitated by a significant increase in the availability of data recording methods and a decrease in the cost of computing power. When designing closed life support systems, data continue to be scarce and expensive. Theoretical approaches based on detailed descriptions of the components of complex systems can predict useful approaches to data preprocessing. Mathematical models, seeking to describe a complex system in a minimally complex way, transform an array of experimental data into a form convenient not only for analysis, but also for perception by a human operator. In addition, mathematical models help solve problems that are still relevant today. We illustrate these points using the example of controlling the assimilation quotient (AQ) of a mixed culture of two algae.

Can we learn anything from the early experience of prototyping closed circuit life support systems (CLSS)? The history of creating closed life support systems goes back more than half a century. Due to the revival of interest in creating bases on the Moon and Mars in the last decade, the relevance of this area of work has increased markedly (Keller et al., 2021; Liu et al., 2021). Since at the initial stages some prototypes were created and studied in detail, later rejected for various reasons, there is a desire to study the experience of these works for possible use in modern projects. Most modern publications persistently propose universal approaches to creating individual life support modules and testing the system (Heinicke, Verseux, 2023; Metelli et al., 2023). Can old approaches be useful for new projects? Also, there is a temptation to compare the approaches used then with those common now, in particular with big data analysis methods.

It is convenient to conduct a similar mental experiment at a preliminary stage for a system that has a fairly detailed theoretical description. In our case, this is a system for co-cultivating two algae (*Chlorella vulgaris* and *Spirulina platensis*) used as an oxygen generator for life support systems (LSS). The idea of using algae to create life support systems is still relevant (Häder, 2020; Fahrion et al., 2021; Matula et al., 2021; Keller et al., 2023). In particular, *Chlorella vulgaris* and *Spirulina platensis* are still actively considered as promising species for this task (Helisch et al., 2020; Cycil et al., 2021; Matula, Nabity, 2021; Matula et al., 2021). We cannot confidently say that all the authors of these works are sincerely convinced of the future role of microalgae in LSS. We believe that higher plants are more promising for solving the problem of providing humans with oxygen and food. However, we, like perhaps many of the authors listed, consider microalgae to be a successful teaching aid. Due to a number of advantages, the cultivation of microalgae is a good model object. For example, in relatively recent literature one can find works devoted to the management of microalgae monocultures (Hu et al., 2008, 2012, 2014), which demonstrate the effectiveness of various management methods. That is, the theoretical work on managing microalgae cultivation in a series of three articles is methodological in nature. We see an opportunity to complement this series of articles by turning to the analysis of the model of forty years ago. As part of the work to create closed life support systems, in 1982, a model for managing a mixed flow-through culture of two algae was created (Bolsunovskiy, Degermendzhi, 1982).

The use of algae as the only autotrophs in the life support system allows us to apply a convenient simplification to reasoning about the stoichiometry of oxygen reduction and carbon dioxide sequestration in an algal cultivator. To a first approximation, we can assume that all carbon dioxide is released by the human body in the oxidation reactions of carbohydrates and fats. This assumption is based on the fact that the use of amino acids by the human body as a significant source of energy is possible with an unbalanced diet, excessive physical activity, or with certain chronic diseases. Having ruled out these three possibilities, we will assume that amino acids make a negligible contribution to respiration. Carbohydrates and fats are the main sources of energy for the human body and the main products of algae biosynthesis.

Another convenient simplification would be to ignore the synthesis of amino acids by algae. Unfortunately, the biomass composition of both algae indicates that proteins are present in large quantities. However, we can allow a first approximation, which should be followed by adjustments to the model if it is necessary to close the nitrogen exchange. That is, to a first approximation, as much as a person oxidizes carbohydrates and fats, the same amount of carbohydrates and fats should be synthesized by algae to bind excess carbon dioxide and regenerate the oxygen used by the person. The use of higher plants would not allow us to resort to such a simple first approximation, since in addition to carbohydrates, fats and proteins, the composition of higher plants contains lignin in noticeable quantities, which differs significantly in stoichiometry from both carbohydrates and fats.

Depending on the diet and level of physical activity, the human body can use different substrates to obtain energy. With sufficient oxygen availability, the main source of energy is the oxidation of fatty acids in mitochondria. When there is a lack of oxygen, the human body prefers carbohydrates as the main source of energy. Thus, the ratio of carbon dioxide emitted by a person and oxygen absorbed can vary from almost 0.7 (oxidation of fats) to 1.0 (oxidation of carbohydrates). For a person, there is even a possibility of a short-term excess of the respiratory index of 1.0 as a result of intense physical activity (acidosis with loss of bicarbonates) and even a long-term excess under the condition of carbohydrate nutrition and an increase in body weight with the accumulation of fat. Unlike humans, algae, on average, maintain a relatively constant composition during their life cycle. Since no synchronization or fluctuations in abundance were observed in the analyzed flow culture, it is possible to use average values of assimilation indices for each of the two algae species.

Assimilation indices reflect the stoichiometric proportion in which the bound carbon dioxide molecules relate to the produced oxygen molecules. Since we agreed to describe the entire metabolism as a first approximation by the balance of fats and carbohydrates, we will leave outside the scope of this article the study of the possibility of shifting the assimilation index of algae by variations in nitrogen nutrition (Belyanin et al., 1980, p. 114–117). We will consider the situation with nitrogen nutrition to be stable and assume that the assimilation index of a system of two algae can vary within the limits indicated in the literature. The metabolic constancy of autotrophs

and human metabolic plasticity must somehow be reconciled within the framework of the work of the CLSS. The range of possible total assimilation index of two algae limits the diet and metabolic activity of a person settled in the CLSS. An important assumption will be that we can adhere to the average specified range by rationally managing a person's diet and physical activity. Then, for example, depending on a long-term increase in the level of physical activity, a person's respiratory coefficient may shift, which will require a shift in the assimilation quotient of the life support system. The design of the life support system should allow for the ability to adapt to the needs of human metabolism. In the analyzed model, we will be interested in the possibility of controlling the composition of a mixed algae culture and controlling the total assimilation quotient.

Materials and methods

Assessment of assimilation indices of a mixed culture of two algae. In order to imagine in more detail the processes of gas exchange in the system under study, we will use the gross formulas of the biomass of chlorella ($C_{6.0}H_{9.7}O_{2.635}N_{0.937}$) (Belyanin et al., 1980, p. 111) and spirulina ($C_{6.0}H_{10.84}O_{2.06}N_{0.87}$) (Belyanin et al., 1980, p. 116). Since the system is considered not closed in nitrogen at the first stage, it is possible to simplify the formulas by considering that the main form of nitrogen absorption by algae is urea or ammonium ions, and also by removing oxygen in the form of water from the formulas. We obtain the residue in the form ($C_{6.0}H_{1.6}$) for chlorella and ($C_{6.0}H_{4.11}$) for spirulina. So, it turns out that the synthesis of chlorella and spirulina biomass allows one absorbed liter of carbon dioxide to release 1.13 liters and 1.3425 liters of oxygen, respectively, which corresponds to the assimilation quotients AQ = 0.885 for chlorella and AQ = 0.745 for spirulina.

The assimilation index of a mixed culture can be easily obtained from the mass ratios of algae in the culture:

$$AQ = X \cdot 0.885 + (1 - X) \cdot 0.745,$$

where X is the proportion of spirulina in the culture. So, for the initially obtained stable mixed culture $X = 0.6$ and $AQ = 0.6 \cdot 0.885 + 0.4 \cdot 0.745 = 0.829$. Controlling the composition of a mixed culture makes it possible to obtain an AQ value ranging from 0.745 (*Spirulina* monoculture) to 0.885 (*Chlorella* monoculture).

Mathematical model. In order to predict the stationary state of algae populations in a flow cultivator, a mathematical model that summarizes information about the influence of control factors on a system of two species is needed. It is precisely this model of a flow cultivator with two algae that was built in (Bolsunovskiy, Degermendzhi, 1982). The model describes the coexistence of two species competing for a limiting substrate. The limiting substrate in this case is the luminous flux. In the model, there is a region of illumination parameters in which two species stably coexist; in addition, there are areas of dominance for each species, when the competing species is forced out. Of course, there is also a range of parameters that does not allow any of the species to reproduce; they are simply washed out of the cultivator with a given flow and insufficient lighting. The flow of the substance in the cultivator was stabilized by recording the absorption of chlorophyll at a wavelength of 680 nm, that is, the system maintained a constant optical density of the medium. The system can be controlled by adjusting the flow rate (that is, the optical density of the medium in the cultivator) and the light intensity. The model does not take into account the photoinhibition of spirulina growth at high light intensity, as well as the effects of metabolic inhibition at high population densities. The mathematical part of the model was obtained as a result of a quantitative description of experiments (Belyanin, Bolsunovskiy, 1980) using differential equations with the subsequent linearization procedure (Bolsunovskiy, Degermendzhi, 1982).

The model is a system of two differential equations, each of which reflects the population dynamics of one alga. The equations look like:

$$X_1' = (\mu_1 - D_f)X_1;\ \mu_1 = a_1 E/(b_1 + E),$$
$$X_2' = (\mu_2 - D_f)X_1\ ;\ \mu_2 = a_2 E/(b_2 + E),$$
$$E = E_0(1 - \gamma_1 X_1 - \gamma_1 X_2),$$
$$D_f = \mu_1 X_1 + \mu_2 X_2.$$

E is average illumination, taking into account the absorption of light by algae cultures. E was obtained after expansion into a Taylor series and discarding nonlinear terms, taking into account the low optical density of the mixed culture. D_f is the flow rate, which in further analysis is replaced by the optical density of the culture as an experimentally measured value. The equations reflect competition for light as a substrate. This substrate, as is known from experimental data on monocultures, is absorbed according to the Michaelis–Menten equation. Specific growth curves in monocultures demonstrate that *Spirulina* is more efficient at light uptake at low light levels, while *Chlorella* is more efficient at high light levels (Fig. 1).

In the parameter ranges characteristic of a stable joint culture of two algae (low population density and low light flux), the model should give the smallest discrepancy with experimental data. To change the ratio of species in the cultivator under these conditions, a small change in the lighting regime or a corresponding change in the flow is sufficient. Long-term increases and decreases in oxygen demand in the CLSS can be compensated by appropriately scaling the cultivator.

Results

The first impression is that the culture of two practically non-interacting species, when competing for a single common substrate, should lead to a stable state when one species dominates and the other species is displaced. It turns out that it is possible to understand why coexistence occurs by carefully analyzing the interaction of species with the substrate in a monoculture. *Chlorella* not only does better in high light, but it also creates some advantage for *Spirulina* in a mixed culture compared to a monoculture. In fact, in the presence of chlorella, spirulina can exist in areas of higher light. Chlorella "shadows" spirulina, creating more comfortable conditions for it. A more detailed analysis of the biology of these species made it possible to identify adaptations to high and low light levels, as well as adaptation to different spectral ranges (Bolsunovskiy, Degermendzhi, 1982). But even without taking into account this adaptability to different parts of the spectrum and, in fact, using the material of experiments with monocultures, it is possible to obtain non-trivial dynamics in

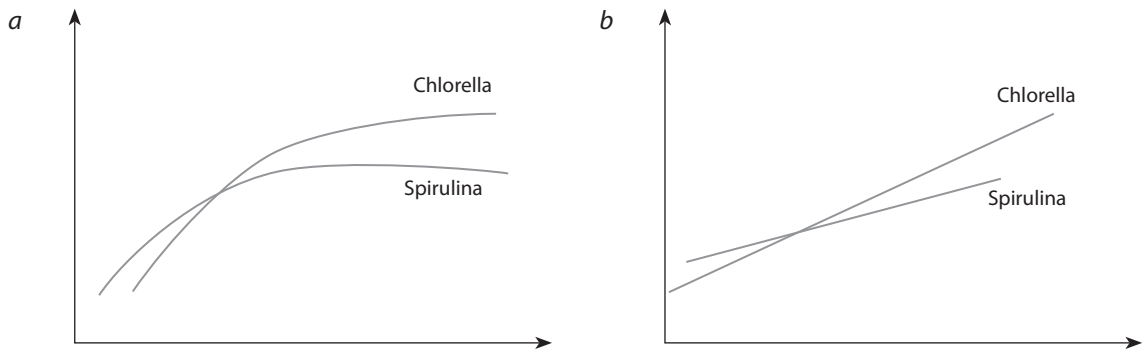


Fig. 1. Specific growth rate depending on illumination of monocultures of chlorella and spirulina (*a*). Representing the same data in inverse coordinates (*b*) shows good agreement with the Michaelis–Menten equation.

the mixed culture model. A mathematical model helps move from qualitative explanation to quantitative predictions.

The model allows us to obtain an area of sustainable coexistence of two species in a continuous culture. Graphically, the area is presented on a plane in the coordinates of illumination (E_0) and the optical density of the crop at a wavelength of 680 nm (C), reflecting the flow rate in the cultivator (Fig. 2).

It is necessary to note that extrapolation of the model results to the area of high illumination and high density of culture is undesirable, since in this area the effect of factors not taken into account in the modeling has been experimentally shown (Belyanin et al., 1980, p. 32–48).

The model allows you to calculate stationary concentrations of components, that is, the population density of individual species:

$$\alpha_1 X_1 = 2K_1K_2(E/E_0 - 1 + C/2K_2)/(K_1 - K_2),$$

$$\alpha_2 X_2 = 2K_1K_2(-E/E_0 + 1 - C/2K_1)/(K_1 - K_2).$$

The steady-state concentrations of each algae species are designated in the model as X_1 and X_2. Since we assume that each alga is characterized by a strictly defined composition and AQ values, the total AQ is a simple superposition: $AQ = (X_1 \cdot AQ_1 + X_2 \cdot AQ_2)/(X_1 + X_2)$. That is, the $X_1/X_2 = \text{const}$ curves will simultaneously be curves with a constant value of the total AQ. It seems paradoxical that all these curves intersect at one point, but the paradox is resolved simply because at this point $X_1 = X_2 = 0$. That is, it does not matter what the ratio of oxygen produced to carbon dioxide absorbed is if the rate of photosynthesis drops to zero.

For the task of controlling the gas composition in a gas-processing facility, it is important to determine where the relation $X_1/X_2 = \text{const}$ is satisfied. The mathematical model was intended to qualitatively explain the observed phenomenon, namely the stable coexistence of two species. One cannot expect an accurate prediction of equilibrium positions over the entire region of existence of the system, but the model can provide a good first approximation for solving such a problem in practice.

Such an approximate algorithm for searching for the equilibrium state of the system will serve as a "rough tuning knob." A more accurate selection of parameters can be carried out experimentally.

In order to understand how a model can be used to analyze experimental data, let's imagine that there are data, but there

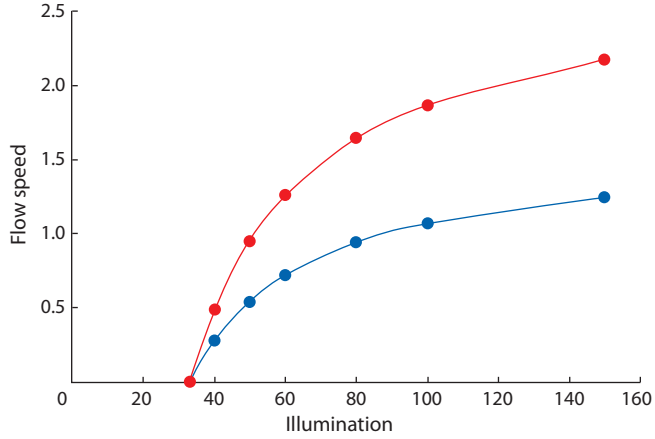


Fig. 2. The area of existence of a stable culture of two algae is limited by two curves on the Illumination/Flow Speed plane.

are no theoretical ideas about how the system functions. A pragmatic approach would be to search for the transformation of curves that limit the region of existence into straight lines in new coordinates. Then all straight lines on this plane passing through the intersection point and lying in the area of existence of a mixed culture could be taken as $X_1/X_2 = \text{const}$. For example, for a given type of curve, an approximation could be a transformation of the form

$$C(E) = K \cdot \ln(E) - \text{const},$$

where K and const would be selected using the least squares method.

Figure 3 shows the results of the inverse transformation of the $E = \exp(C/K + \text{const})$ graphs. It can be noted that after the transformation the points are well approximated by a straight line.

All possible stable equilibrium positions of the system that allow the coexistence of two species can, after such a transformation, be represented by a bunch of straight lines passing through one point. For each such line we can take AQ = const. Since AQ is obtained by a simple superposition of the assimilation indices of two algae, it is natural to assume that on a plane where data on monocultures are represented by straight lines, data on a mixed culture will also be represented by straight lines.

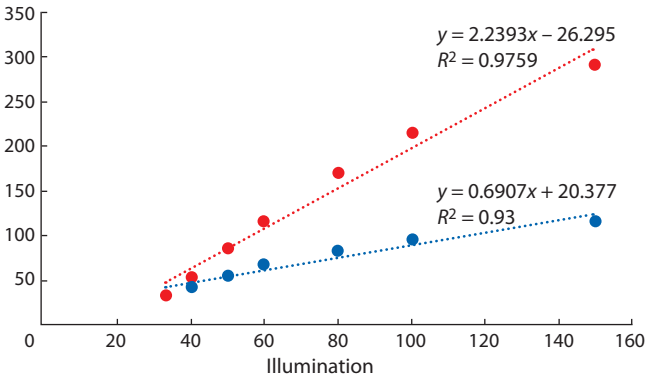


Fig. 3. The result of the empirical selection of a transformation that "straightens" the data graphs in new coordinates.

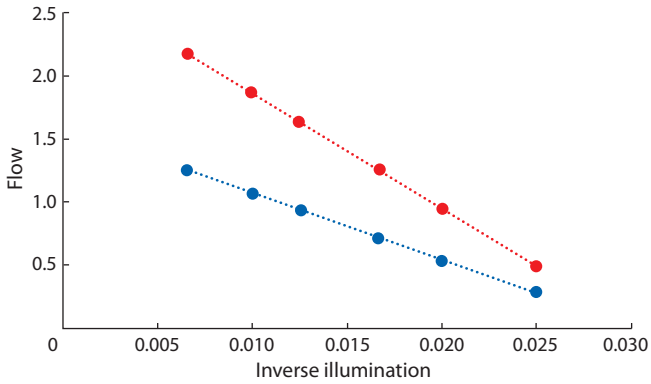


Fig. 4. As follows from the model, to search for equilibrium positions in a mixed culture, it is convenient to present data in coordinates (Inverse Illumination/Flow).

It is worth paying attention to two obvious facts: (1) the chosen approximation is sensitive to the area in which the experimental data are collected; (2) the approximation produces a systematic error, underestimating the results at average illumination and overestimating them in the areas of low and high illumination.

Now let's compare this approach with the one that follows from knowing the exact solution of the model. An exact approximation of the model solution will be given by transformation to coordinates ($1/E_0$; C). Exact solutions are then converted to straight lines (Fig. 4). All points that obey the relations X_1/X_2 = const will also lie on straight lines passing through the common intersection point. It is this approximation that can be recommended for further use in processing experimental data as a first approximation.

Let's imagine a situation where we have experimental data obtained under modern conditions. Let's say a stationary state has been established in the cultivator. In the experiment, you can control the flow rate and illumination. Using gas analysis, you can obtain the AQ value for a stationary case, and then calculate the ratio of species in the culture. One can also imagine a direct measurement of the species ratio. Using modern methods, for example, flow cytometry, it is possible to automatically obtain data on the steady-state X_1/X_2 ratio. All this data can be used to restore the parameters of calibration graphs of the form X_1/X_2 = const. That is, the theory helps to choose a data preprocessing procedure for further analysis, for example, using methods of mathematical statistics, or artificial neural networks, or even in the form of graphical constructions. Moreover, the theory was obtained based primarily on data on the specific growth rate of algae in monocultures. Based on data on monocultures, empirical methods simply cannot predict the relationships in a mixed culture, so empirical methods, which include all modern "methods of big data analysis," will require not only large, but also rather hard-to-access data.

How can we now determine the position of a straight line with a given ratio X_1/X_2? The bottom graph is the optical density of the spirulina monoculture, the top graph is the optical density of the chlorella monoculture. In order to find a position with a given ratio X_1/X_2 at a given level of illumination, it is necessary to divide the vertical segment connecting the lower and upper straight lines in the ratio X_1/X_2. The stability of the solution of the mathematical model guarantees that subsequent experimental refinement of the equilibrium position will be small. Empirical methods currently do not provide insight into the stability of the predictions obtained with their help.

Conclusion

When creating complex biotechnological systems, fairly simple and visual mathematical models can be a good addition to modern methods of data analysis. If experimental data are difficult to access, the only way to predict the behavior of the system is to create an adequate mathematical model. In addition, in the case of closed life support systems, the ability to understand the structure of the system on the part of the human operator, as a rule, the occupant of this system, is important. The simpler and more obvious the mechanisms incorporated into the design of the life support system, the higher its reliability will be.

References

Belyanin V.N., Bolsunovskiy A.Ya. Regulation of species range in a two-component algae community in an experiment. In: Parametric Control of Microalgal Biosynthesis. Novosibirsk: Nauka Publ., 1980;72-80 (in Russian)

Belyanin V.N., Sydko F.Ya., Trinkenschu A.P. Energetics of Photosynthesizing Plant Culture. Novosibirsk: Nauka Publ., 1980 (in Russian)

Bolsunovskiy A.Ya., Degermendzhi A.G. Study of the photosynthetic mechanism of coexistence of species in a mixed continuous-flow chlorella-spirulina culture. In: Issues of Controlling the Biosynthesis in Lower Plants. Novosibirsk: Nauka Publ., 1982;99-116 (in Russian)

Cycil L.M., Hausrath E.M., Ming D.W., Adcock C.T., Raymond J., Remias D., Ruemmele W.P. Investigating the growth of algae under low atmospheric pressures for potential food and oxygen production on Mars. *Front. Microbiol.* 2021;12:733244. DOI 10.3389/fmicb.2021.733244

Fahrion J., Mastroleo F., Dussap C.-G., Leys N. Use of photobioreactors in regenerative life support systems for human space exploration. *Front. Microbiol.* 2021;12:699525. DOI 10.3389/fmicb.2021.699525

Häder D. On the way to Mars-flagellated algae in bioregenerative life support systems under microgravity conditions. *Front. Plant Sci.* 2020;10:1621. DOI 10.3389/fpls.2019.01621

Heinicke C., Verseux C. The MaMBA facility as a testbed for bioregenerative life support systems. *Life Sci. Space Res.* (*Amst.*). 2023; 36:86-89. DOI 10.1016/j.lssr.2022.08.009

Helisch H., Keppler J., Detrell G., Belz S., Ewald R., Fasoulas S., Heyer A.G. High density long-term cultivation of *Chlorella vulgaris* SAG 211-12 in a novel microgravity-capable membrane raceway photobioreactor for future bioregenerative life support in SPACE. *Life Sci. Space Res.* (*Amst.*). 2020;24:91-107. DOI 10.1016/j.lssr.2019.08.001

Hu D., Liu H., Yang C., Hu E. The design and optimization for light-algae bioreactor controller based on Artificial Neural Network-Model Predictive Control. *Acta Astronaut*. 2008;63(7-10):1067-1075. DOI 10.1016/j.actaastro.2008.02.008

Hu D., Li M., Zhou R., Sun Y. Design and optimization of photo bioreactor for O_2 regulation and control by system dynamics and computer simulation. *Bioresour. Technol*. 2012;104:608-615. DOI 10.1016/j.biortech.2011.11.049

Hu D., Li L., Li Y., Li M., Zhang H., Zhao M. Gas equilibrium regulation by closed-loop photo bioreactor built on system dynamics, fuzzy inference system and computer simulation. *Comput. Electron. Agric*. 2014;103:114-121. DOI 10.1016/j.compag.2014.02.002

Keller R.J., Porter W., Goli K., Rosenthal R., Butler N., Jones J.A. Biologically-based and physiochemical life support and in situ resource utilization for exploration of the Solar System – reviewing the current state and defining future development needs. *Life*. 2021; 11(8):844. DOI 10.3390/life11080844

Keller R., Goli K., Porter W., Alrabaa A., Jones J.A. Cyanobacteria and algal-based biological life support system (BLSS) and planetary surface atmospheric revitalizing bioreactor brief concept review. *Life*. 2023;13(3):816. DOI 10.3390/life13030816

Liu H., Yao Z., Fu Y., Feng J. Review of research into bioregenerative life support system(s) which can support humans living in space. *Life Sci. Space Res.* (*Amst.*). 2021;31:113-120. DOI 10.1016/j.lssr.2021.09.003

Matula E.E., Nabity J.A. Effects of stepwise changes in dissolved carbon dioxide concentrations on metabolic activity in *Chlorella* for spaceflight applications. *Life Sci. Space Res.* (*Amst.*). 2021;29:73-84. DOI 10.1016/j.lssr.2021.03.005

Matula E.E., Nabity J.A., McKnight D.M. Supporting simultaneous air revitalization and thermal control in a crewed habitat with temperate *Chlorella vulgaris* and eurythermic antarctic chlorophyta. *Front. Microbiol*. 2021;12:709746. DOI 10.3389/fmicb.2021.709746

Metelli G., Lampazzi E., Pagliarello R., Garegnani M., Nardi L., Calvitti M., Gugliermetti L., Restivo Alessi R., Benvenuto E., Desiderio A. Design of a modular controlled unit for the study of bioprocesses: eowards solutions for Bioregenerative Life Support Systems in space. *Life Sci. Space Res.* (*Amst.*). 2023;36:8-17. DOI 10.1016/j.lssr.2022.10.006

ORCID ID

D.A. Semyonov orcid.org/0000-0002-4993-6358
A.G. Degermendzhi orcid.org/0000-0001-8649-5419

Acknowledgements. The study was supported by the Russian Science Foundation grant No. 23-44-00059, https://rscf.ru/project/23-44-00059/

Conflict of interest. The authors declare no conflict of interest.

Received July 20, 2023. Revised September 18, 2023. Accepted September 19, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):884-889
DOI 10.18699/VJGB-23-102

A phenomenological model of non-genomic variability of luminescent bacterial cells

S.I. Bartsev[1, 2]

[1] Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
[2] Siberian Federal University, Krasnoyarsk, Russia
bartsev@yandex.ru

Abstract. The light emitted by a luminescent bacterium serves as a unique native channel of information regarding the intracellular processes within the individual cell. In the presence of highly sensitive equipment, it is possible to obtain the distribution of bacterial culture cells by the intensity of light emission, which correlates with the amount of luciferase in the cells. When growing on rich media, the luminescence intensity of individual cells of brightly luminous strains of the luminescent bacteria *Photobacterium leiognathi* and *Ph. phosporeum* reaches 10^4–10^5 quanta/s. The signal of such intensity can be registered using sensitive photometric equipment. All experiments were carried out with bacterial clones (genetically homogeneous populations). A typical dynamics of luminous bacterial cells distributions with respect to intensity of light emission at various stages of batch culture growth in a liquid medium was obtained. To describe experimental distributions, a phenomenological model that links the light of a bacterial cell with the history of events at the molecular level was constructed. The proposed phenomenological model with a minimum number of fitting parameters (1.5) provides a satisfactory description of the complex process of formation of cell distributions by luminescence intensity at different stages of bacterial culture growth. This may be an indication that the structure of the model describes some essential processes of the real system. Since in the process of division all cells go through the stage of release of all regulatory molecules from the DNA molecule, the resulting distributions can be attributed not only to luciferase, but also to other proteins of constitutive (and not only) synthesis.
Key words: non-genomic variability; phenomenological model; luminescent bacteria.

For citation: Bartsev S.I. A phenomenological model of non-genomic variability of luminescent bacterial cells. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):884-889. DOI 10.18699/VJGB-23-102

Феноменологическая модель негеномной изменчивости люминесцентных бактериальных клеток

С.И. Барцев[1, 2]

[1] Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
[2] Сибирский федеральный университет, Красноярск, Россия
bartsev@yandex.ru

Аннотация. Свет, испускаемый люминесцентными бактериями, может служить уникальным природным каналом передачи информации о процессах внутри отдельной клетки. При наличии высокочувствительного оборудования можно получить распределение клеток бактериальной культуры по интенсивности свечения, которая коррелирует с количеством люциферазы в клетках. При выращивании на богатых питательных средах интенсивность свечения отдельных клеток ярко светящихся штаммов люминесцентных бактерий *Photobacterium leiognathi* и *Ph. phosporeum* достигает 10^4–10^5 квантов/с. Сигнал такой интенсивности может быть зарегистрирован с помощью чувствительного фотометрического оборудования. Все эксперименты проводились с бактериальными клонами – генетически однородными популяциями. Получена типичная динамика распределения светящихся бактериальных клеток по интенсивности свечения на различных стадиях периодического выращивания культуры в жидкой среде. Для описания экспериментальных распределений была построена феноменологическая модель, которая связывает излучение бактериальной клетки с историей событий на молекулярном уровне. Предложенная феноменологическая модель с минимальным числом подстроечных параметров (1.5) обеспечивает удовлетворительное описание сложного процесса формирования распределения клеток по интенсивности свечения на разных стадиях роста бактериальной культуры. Это может свидетельствовать о том, что структура модели описывает некоторые существенные процессы реальной системы. Поскольку в процессе деления все клетки проходят стадию отсоединения всех регуляторных молекул от молекулы ДНК, результирующие распределения можно отнести не только к люциферазе, но и к другим белкам конститутивного (и не только) синтеза.
Ключевые слова: негеномная изменчивость; феноменологическая модель; люминесцентные бактерии.

Introduction

The heterogeneity of isogenic bacterial populations, or, in other words, non-genomic variability of cells, is increasingly attracting the attention of researchers. This is partly due to the development of methods for tracking individual cell parameters, down to the dynamics of protein synthesis during the cell cycle (Taheri-Araghi et al., 2015; Andryukov et al., 2021). On the other hand, understanding the mechanisms or causes of phenotypic differences of cells from an isogenic population is important both for the formation of fundamental concepts of intracellular processes organization and for increasing the efficiency of solving practical problems in medicine and biotechnology.

The cell cycle is a potentially significant source of non-genomic variability. During the cell cycle, the protein abundance in the cell undergoes two-fold changes. In the case of an asynchronous population, these changes can contribute significantly to phenotypic variability. However, another possible source of heterogeneity is related to the cell cycle. It has been shown quite a long time ago (Shkolnik, 1989) that the widely used allometric dependences (when different variables N_i are related by relations of the form $N_i = \alpha_i N_1^{\beta_i}$), when describing growth curves, lead to a contradiction with observations. So in the case of an allometric growth model, a cell dies after a small number of generations due to the fact that certain substances abundance approaches zero. Then a phenomenological trigger model combining allometric growth with switches was proposed. According to the model, the passage of a cell through various phases of the cell cycle is accompanied by sharp changes in the allometric ratios of growth variables. There are certain combinations of parameters that can be conditionally associated with multidimensional switching surfaces – the boundaries of cellular phases – from cell birth to division. When passing the next boundary, the rates of change in cellular variables switch. This model was further developed (Zinovyev et al., 2022) and demonstrated strong agreement with experimental data.

According to this model, switching should occur in a certain sequence and in a fairly uniform manner, but for a non-synchronous culture such switching can make a significant contribution to the variability of phenotypic traits. However, it should be noted that this model was compared with data on the dynamics of variable eukaryotic cells and it is possible that in bacterial cells the limitations of allometric growth are overcome in another way.

Thus, experimental observations of protein synthesis inside bacterial cells (Kiviet et al., 2014) show that the activation of particular protein synthesis occurs without pronounced patterns. Another paper on the topic (Walker et al., 2016) notes that the contribution of the bacterial cell cycle to expression noise consists of two parts: a deterministic fluctuation synchronous with the cell cycle and a stochastic component caused by variable timing of gene replication. It was shown earlier (Taniguchi et al., 2010) that proteins with strong expression have a coefficient of variation of ~30 %, which indicates an "external" factor not associated with fluctuations in the abundance of a small number of molecules.

Fluorescence microscopy is primarily used to monitor protein synthesis at the single-cell scale, which is essential for studying non-genomic variation. However, it is noted that with the current level of device sensitivity stimulating light has a negative effect on the physiological state of cells (Taheri-Araghi et al., 2015).

A unique alternative to fluorescence microscopy is the use of luminescence of luminescent bacteria (Deryabin, 2009) as a channel of information about the state of intracellular processes (Berzhanskaya et al., 1975; Bartsev, Gitelzon, 1985). The uniqueness of luminescence lies in the fact that the cell emits light while in its native state, which significantly reduces the probability of artifacts. Moreover, since the intensity of cell luminescence depends both on the abundance of luciferase and on the presence of substrates for the luciferase reaction, the luminescence of a bacterium is a kind of multiplexer – information from different input channels can be transmitted through one output channel – about the expression of the luciferase operon, on the one hand, and the state of the cell's energy metabolism, on the other.

The goal of the work is to assess the degree of variability of individual bacterial cells regarding luminescence intensity at different stages of development of batch culture of bacteria, and to test the simplest possible approach to the mathematical description of this variability.

Experiment description

When growing on rich media, the luminescence intensity of individual cells of brightly luminous strains of luminescent bacteria *Photobacterium leiognathi* and *Ph. phosporeum* reaches 10^4–10^5 quanta/s. Such signal can be registered using sensitive photometric equipment. The strains used did not demonstrate the typical quorum effect (Brodl et al., 2018) and an increase in their luminescence was observed from the beginning of culture growth.

Without delving into the details of the experimental setup, which operates in the photon counting mode, and the routine for measuring the distribution of bacterial cells according to luminescence intensity (Bartsev, Shenderov, 1985), let us proceed to the description of the results. It should be noted that all experiments were carried out with bacterial clones (genetically homogeneous populations).

During the registration of distributions, the bacteria were in a medium containing only glucose as an energy substrate, i. e. bacterial growth was stopped and the luciferase abundance during the measurement can be considered unchanged. At least, control experiments showed that over a typical period of time the luminescence intensity of individual bacterial cells did not undergo noticeable changes.

A typical view of luminous bacteria distribution at various stages of batch culture growth in a liquid medium is shown in Figure 1.

An immediate question arises regarding the potential mechanism behind the observed variation in the phenotypic trait. The simplest explanation for the observed variability can be suggested immediately – the intensity of the emission is determined by the variability of the bacterial cell volumes. However, direct measurements of cell volume variation in *B. subtilis* and *E. coli* showed that the coefficient of variation (CV) of cell volume is ~23 % (van Heerden et al., 2017), while the average CV of bacterial luminescence intensity

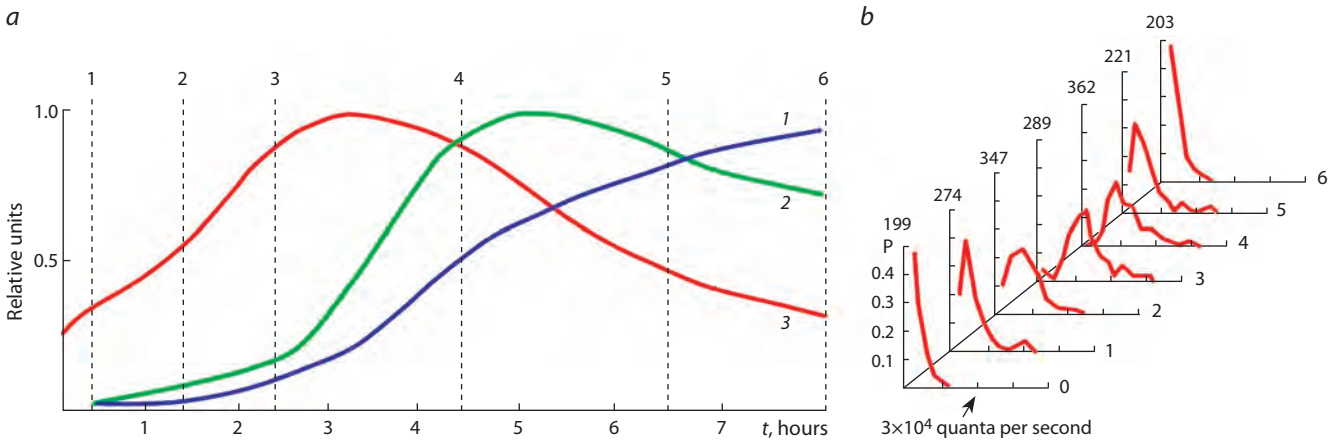


Fig. 1. Dynamics of luminescent bacteria culture parameters (*a*) and cell distributions by luminescence intensity (*b*).

Curves of culture parameters are given in relative units: *1* – optical density; *2* – culture luminescence intensity; *3* – the average intensity of a single cell. The dashed lines indicate sampling times, and their numbers correspond to the numbers of distributions.

is ~50 % and can exceed 70 %. Therefore, there is an additional factor that provides a significant variability in cell luminescence.

On possible causes of non-genomic variability

Under normal growth conditions, the luminescence intensity of a bacterial cell is determined by the abundance of luciferase, the enzyme responsible for catalyzing the luminescent reaction, as well as a set of enzymes that supply the necessary substrates for this reaction (Brodl et al., 2018). Proteins involved in bacterial bioluminescence, notably, LuxCDABEG, are encoded by the lux operon and are highly conserved among different bacterial strains. The *luxA* and *luxB* genes encode a heterodimeric luciferase; the *luxCs*, *luxDs*, and *luxE* gene products are components of the fatty acid reductase complex; and *luxG* encodes flavin reductase.

It is natural to assume that in the presence of an energy substrate, as was the case in the experiments performed, the intensity of bacterial luminescence is determined primarily by the expression of the luciferase operon. Other factors, such as the contribution of uneven distribution of protein, mRNA and ribosomes during division, variability in the amount of mRNA due to the small number of molecules, the transition of genes from active to passive state due to reversible binding of a transcription factor, conformation of the DNA molecule that prevents binding RNA polymerases show less variability (Paulsson, 2004; Schwabe, Bruggeman, 2014; Kuwahara et al., 2015; van Heerden et al., 2017; Dessalles et al., 2020) than observed in the experiment. In addition, the resulting cell distributions by protein amount give a distribution close to normal, while asymmetric distributions were observed in the experiment. In addition to this, these distributions demonstrated characteristic dynamics during the development of the enrichment culture, and an adequate model for the formation of distributions of luminescent bacteria by luminescence intensity should, at least qualitatively, reproduce the experimental dynamics.

With a large number of molecules, which is the case for luciferase, fluctuations in its amount between daughter cells are determined by fluctuations in the uneven volumes of daughter cells, which cannot explain the observed CV value. At the same time, it was shown (Taniguchi et al., 2010) that proteins with strong expression have a coefficient of variation of ~30 %, which indicates an "external" factor not associated with fluctuations in a small number of molecules.

Mathematical model derivation

Without delving into the details of the processes of transcription and translation, let us consider a possible phenomenological stochastic mechanism for generating significant variability in the amount of luciferase in cells. The amount of luciferase in a cell of age $\tau - z(\tau)$ is the sum of the amount of luciferase received by the cell after division (x) and the amount of luciferase accumulated by age $\tau - y(\tau)$:

$$z(\tau) = x + y(\tau). \quad (1)$$

Immediately after division, when $\tau = 0$, the cell contains only the luciferase produced in the previous cell cycle. Let $f(x)$ be the distribution of cells of a narrow age interval according to the amount of luciferase obtained during division, which does not change throughout the entire cell cycle. The form of this distribution is not known and must be obtained by solving the model equation.

Type of cells distribution from a narrow age interval according to the amount of luciferase synthesized and accumulated by age $\tau - P(y, \tau)$ can be obtained from the following considerations. For the sake of simplicity, let's assume that luciferase synthesis begins immediately after cell division, closely associated with the release of DNA from all transcription factors (in our case, the luciferase gene repressor), proceeds at a constant rate, and stops after binding the repressor to the operator.

Let's assume that τ' is the moment when the repressor binds to the operator. Then the amount of luciferase synthesized by time τ is described by the following expression:

$$y(\tau) = \alpha \int_0^{\tau} \theta(\tau' - \eta) d\eta, \quad (2)$$

where α is the rate of enzyme synthesis; θ is the Heaviside step function.

Since $y(\tau)$ is also a function of the random variable τ', distribution $P(y, \tau)$ is described by the following expression:

$$P(y,\tau)=\int_0^{\tau} g(\tau')\delta(y-\alpha\tau')\,d\tau'+\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau', \quad (3)$$

where $g(\tau')$ is the distribution describing the proportion of the cell population in which the binding of the repressor to the operator occurred in the interval $[\tau', \tau'+d\tau']$; $\delta(x)$ is the Dirac delta function.

This integral is split into two integrals with integration limits $[0, \tau)$ and $[\tau, \infty)$, and the cells in which the binding of the repressor to the operator occurred by the age τ ($\tau'<\tau$) fall into the first integral, the rest ($\tau'\geq\tau$) fall into in the second. Let's do some calculations:

$$P(y,\tau)=\int_0^{\tau} g(\tau')\,\delta\Big(y-\alpha\int_0^{\tau'} d\eta\Big)\,d\tau'+\int_{\tau}^{\infty} g(\tau')\,\delta\Big(y-\alpha\int_0^{\tau} d\eta\Big)\,d\tau',$$

$$P(y,\tau)=\int_0^{\infty} g(\tau')\,\delta\Big[y-\alpha\int_0^{\tau}\theta(\tau'-\eta)\,d\eta\Big]\,d\tau',$$

$$P(y,\tau)=\frac{1}{\alpha}\,g\Big(\frac{y}{\alpha}\Big)\theta(\alpha\tau-y)+\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'.$$

Since the total amount of luciferase in a cell ($z(\tau)$) is the sum of independent random variables x and y, then the distribution of cells in a narrow time interval of age τ by the total amount of luciferase has the following form:

$$L(z,\tau)=\int_0^{\infty}\int_0^{\infty} f(x)P(y,\tau)\,\delta(z-x-y)\,dx\,dy,$$

$$L(z,\tau)=\int_0^{\infty} f(z-y)\,P(y,\tau)\,dy,$$

$$L(z,\tau)=\int_0^{\infty} f(z-y)\frac{1}{\alpha}\,g\Big(\frac{y}{\alpha}\Big)\theta\Big(\tau-\frac{y}{\alpha}\Big)dy+$$

$$+\int_0^{\infty} f(z-y)\,\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'\,dy.$$

By changing the variables $\tau' = y/\alpha$ we get:

$$L(z,\tau)=\int_0^{\tau} f(z-\alpha\tau')\,g(\tau')\,d\tau'+f(z-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'. \quad (4)$$

As a result, an expression for the distribution of cells by the amount of luciferase for a narrow age range of age τ was obtained. In order to obtain the equations for the distribution function $f(x)$ and the expression for $\Phi(z)$ – the distribution function of the cell population by the amount of luciferase, it is necessary to know the age structure of the population.

The form of cells distribution by age $\Psi(\tau)$ is obtained from the equation (Romanovsky et al., 1984):

$$\frac{\partial n}{\partial t}+\frac{\partial n}{\partial \tau}=-\omega(\tau)n,$$

where $n(t, \tau)d\tau$ is the number of cells of age in the interval $[\tau, \tau+d\tau]$ at the moment t; $\omega(\tau)$ is the rate of cell loss from a given age interval due to division.

Let us consider the case of a stationary age distribution of bacteria, i.e. $n(t, \tau)/N(t)$, is fixed, but the total number of cells $N(t)$ increases. In the case of a stationary distribution, the specific growth rate of cells number in a given age interval is equal to the specific population growth rate:

$$\frac{\partial n(t,\tau)}{\partial t}=\mu n(t,\tau). \quad (5)$$

Dividing this equation by $N(t)$ we get the equation for frequencies:

$$\frac{\partial \Psi}{\partial \tau}=-[\omega(\tau)+\mu]\Psi, \quad \Psi(\tau)=\frac{n(t,\tau)}{N(t)}.$$

For simplicity, we set the division rate as a step function (Romanovsky et al., 1984, p. 88):

$$\omega(\tau)=C\Theta(\tau-\tau_1)=\begin{cases}0, & \tau<\tau_1\\ C, & \tau\geq\tau_1\end{cases},$$

then the distribution density of dividing cells looks like:

$$\Omega(\tau)=\begin{cases}0 & ,\ \tau<\tau_1\\ Ce^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases},$$

where C is the intensity of cell division events. And as a result:

$$\Psi(\tau)=\begin{cases}\Psi_0 e^{-\mu\tau} & ,\ \tau<\tau_1\\ \Psi_0 e^{-\mu\tau}e^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases}.$$

It remains to determine the form of the function $g(\tau)$. Assumptions about the constant amount of the repressor in the cell and the irreversibility of its binding to the operator allow us to represent the distribution of cells over the time that elapsed from replication (division) to the moment of binding the repressor to the operator in the form of an exponential distribution:

$$g(\tau)=Ae^{-A\tau},$$

where A is the intensity of events.

As a result of all substitutions, we obtain a model for the distribution of luciferase over the cells of the bacterial culture:

$$f\Big(\frac{z}{2}\Big)=2\int_0^{\infty}\Omega(\tau)\,d\tau\left[\int_0^{\tau} f(z-\alpha\tau')Ae^{-A\tau'}d\tau'+2f(z-\alpha\tau)e^{-A\tau}\right],$$

$$\Phi(z)=\int_0^{\infty}\Psi(\tau)\,d\tau\left[\int_0^{\tau} f(z-\alpha\tau')Ae^{-A\tau'}d\tau'+f(z-\alpha\tau')e^{-A\tau}\right],$$

where

$$\Omega(\tau)=\begin{cases}0 & ,\ \tau<\tau_1,\\ Ce^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases}\quad \Psi(\tau)=\begin{cases}\Psi_0 e^{-\mu\tau} & ,\ \tau<\tau_1\\ \Psi_0 e^{-\mu\tau}e^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases},$$

and where $f(z)$ is the density of distribution of cells from a narrow age interval according to the amount of luciferase obtained during division; $\Phi(z)$ is the density of cell distribution according to the intracellular amount of luciferase; $\Psi(\tau)$ is distribution density of culture cells by age; $\Omega(\tau)$ is distribution density of dividing cells; A is the intensity of binding the repressor to the operator; α is the rate of luciferase synthesis; C is the intensity of cell division events; τ_1 is the minimum age of the beginning of cell division τ.

Computer simulation

If the resulting equations cannot be solved analytically, then successive approximations are used. But first the values of the model parameters need to be chosen. Note that if the intensity of the repressor binding the activator (parameter A) is equal to zero, then constitutive protein synthesis throughout the entire cell cycle takes place. It is natural to compare this synthesis with the growth of cell volume.

That is, the parameters C and τ_1 can be determined from other independent distributions (van Heerden et al., 2017), assuming that the coefficients of variation of distributions by volume in luminescent bacteria and other gram-negative bacte-

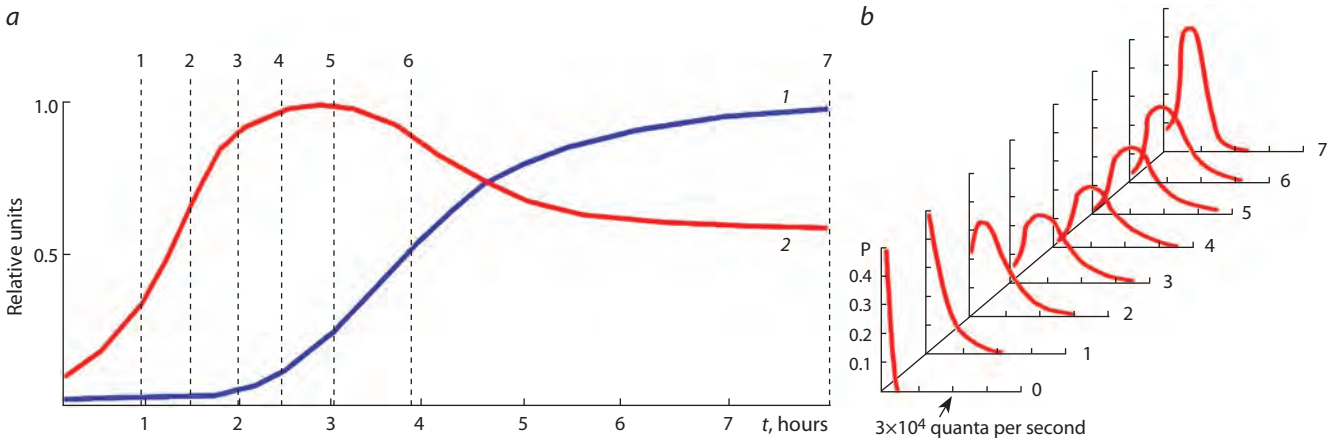


Fig. 2. Model dynamics of luminescent bacteria culture parameters (*a*) and cell distributions by luminescence intensity (*b*).

Curves of culture parameters are given in relative units: *1* – biomass; *2* – the average intensity of a single cell emission. The dashed lines indicate the moments of "sampling", and the numbers correspond to the numbers of the distributions.

ria are close. The coefficient of variation of the model distribution is close to the value of 24 % at $C = 4$ and $\tau_1 = 3/4\ \tau_0$, where τ_0 is the average generation time in the population. These values were used for further simulation. When modeling the dynamics of light intensity distributions during population growth, at the next iteration step the value of the specific growth rate μ was substituted from population growth simulation describing the growth of a real culture.

Thus, as a result, there are only two adjustable parameters, or rather, one and a half – the parameter α (the rate of luciferase synthesis) is, in fact, a scale factor. It shows the relative value of the luminescence intensity, mediated in the experiment by the quantum efficiency of the luciferase itself, the geometry of the recording system that determines the amount of light from a bacterium that hits the photocathode of the photomultiplier, the quantum yield of the photocathode, and the fraction of single-electron pulses cut off by the discriminator at the PMT output.

So to describe the dynamics of distributions obtained in the experiment, the model has one adjustable parameter, *A*, the intensity of repressor-operator binding events. The results of calculations for the most suitable value for describing real distributions, which is $A = 2$, are shown in Figure 2.

When comparing Figures 2 and 1, one can see a quite satisfactory correspondence between them. It is worth noting that this correspondence was obtained with one fitting parameter, which apparently indicates that the proposed model describes something significant in the simulated real system.

It should be noted that luciferase inactivation was not taken into account when deriving the model, which was done to simplify the model; however, it is a common practice (Schwabe, Bruggeman, 2014, p. 306). Palliative inactivation of luciferase can be introduced externally – simply by shifting the distribution points to 0 in proportion to their distance from the origin. In this case, the visual representation of the model would be closer to the experimental data.

However, one property of the model is of interest, which manifested itself in the shift of distributions to 0 at the last stages of population development. By distribution No. 4, the model has almost reached a stationary state and should have remained in it. But since the model takes into account the increase in the duration of the generation time due to the slowdown in culture growth, the established balance between the rate of luciferase synthesis and its distribution between two daughter cells is disturbed.

Since the rate of synthesis of a particular protein is related to the state of basic metabolism, a slowdown in the cell growth rate and accordingly an increase in the generation time leads to a decrease in the rate of luciferase synthesis (decreasing α coefficient). But the intensity of repressor-operator binding events (a physical, energy-independent process) remains the same. However, on the time scale of the cell itself (the unit of measurement is generation time), the rate of luciferase synthesis remained the same, while the intensity of switching events of the luciferase operon increased. Therefore, according to the model, there is a close relationship between the rate of cell growth and the content of luciferase in it, and the higher the rate, the more luciferase is synthesized per cell cycle and vice versa.

The proposed model based on switching off the operon some time after the birth corresponds to the results on the dependence of fluorescent protein expression on cell age (van Heerden et al., 2017, Fig. 4, B, C). It should be noted that the imposition of the age distribution on the expression level curve (Fig. 4, C) was not done entirely correctly by the authors – they have expression even at negative ages (beyond the left border of the age distribution). When bringing the expression level to the age distribution, it would be even more clearly visible, as can be judged by the saturation of the blue area in Fig. 4, B, that the expression level is maximum immediately after the birth of the cell and then decreases with age, which corresponds to the proposed model.

Conclusion

In conclusion, it can be noted that the proposed phenomenological model with a minimum number of adjustable parameters (1.5) satisfactorily describes a rather complex process that takes place during the growth of a bacterial culture. This may be an indication that the structure of the model describes some essential processes of the real system. Since

in the process of division all cells go through the stage of release of all regulatory molecules from the DNA molecule, the resulting distributions can be realized not only in relation to luciferase, but also to other proteins of constitutive (and not only) synthesis.

References

Andryukov B.G., Timchenko N.F., Lyapun I.N., Bynina M.P., Matosova E.V. Heterogeneity in isogenic bacteria populations and modern technologies of cell phenotyping. *J. Microbiol. Epidemiol. Immunobiol.* 2021;98(1):73-83. DOI 10.36233/0372-9311-33 (in Russian)

Bartsev S.I., Gitelson J.I. On the temporary organization of bacterial luminescence. *Studia Biophisica.* 1985;105(3):149-156 (in Russian)

Bartsev S.I., Shenderov A.N. Dynamics of distributions of luminescent bacteria according to the intensity of luminescence in periodic culture. Krasnoyarsk: Preprint Institute of Physics SB AS USSR, 1985 (in Russian)

Berzhanskaya L.Yu., Gitelson J.I., Fish A.M., Chumakova R.I. On the pulsed nature of bacterial bioluminescence. *Doklady Akademii Nauk SSSR.* 1975;222(5):1220-1222 (in Russian)

Brodl E., Winkler A., Macheroux P. Molecular mechanisms of bacterial bioluminescence. *Comput. Struct. Biotechnol. J.* 2018;16:551-564. DOI 10.1016/j.csbj.2018.11.003

Deryabin D.G. Bacterial Bioluminescence: Fundamental and Applied Aspects. Moscow: Nauka Publ., 2009 (in Russian)

Dessalles R., Fromion V., Robert P. Models of protein production along the cell cycle: an investigation of possible sources of noise. *PLoS One.* 2020;15(1):e0226016. DOI 10.1371/journal.pone.0226016

Kiviet D.J., Nghe P., Walker N., Boulineau S., Sunderlikova V., Tans S.J. Stochasticity of metabolism and growth at the single-cell level. *Nature.* 2014;514(7522):376-379. DOI 10.1038/nature13582

Kuwahara H., Arold S.T., Gao X. Beyond initiation-limited translational bursting: the effects of burst size distributions on the stability of gene expression. *Integr. Biol.* 2015;7(12):1622-1632. DOI 10.1039/c5ib00107b

Paulsson J. Summing up the noise in gene net works. *Nature*. 2004; 427(6973):415-418. DOI 10.1038/nature02257

Romanovsky Yu.M., Stepanova N.V., Chernavsky D.S. Mathematical Biophysics. Moscow: Nauka Publ., 1984 (in Russian)

Schwabe A., Bruggeman F.J. Contributions of cell growth and biochemical reactions to nongenetic variability of cells. *Biophys. J.* 2014;107(2):301-313. DOI 10.1016/j.bpj.2014.05.004

Shkolnik E.M. Dynamic models of the cell cycle. In: Bykov V.I. (Ed.) Dynamics of Chemical and Biological Systems. Novosibirsk: Nauka Publ., 1989;230-260 (in Russian)

Taheri-Araghi S., Brown S.D., Sauls J.T., McIntosh D.B., Jun S. Single-cell physiology. *Annu. Rev. Biophys.* 2015;44:123-142. DOI 10.1146/annurev-biophys-060414-034236

Taniguchi Y., Choi P.J., Li G.-W., Chen H., Babu M., Hearn J., Emili A., Xie X.S. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. 2010;329(5991): 533-538. DOI 10.1126/science.1188308

van Heerden J.H., Kempe H., Doerr A., Maarleveld T., Nordholt N., Bruggeman F.J. Statistics and simulation of growth of single bacterial cells: illustrations with *B. subtilis* and *E. coli*. *Sci. Rep.* 2017; 7(1):16094. DOI 10.1038/s41598-017-15895-4

Walker N., Nghe P., Tans S.J. Generation and filtering of gene expression noise by the bacterial cell cycle. *BMC Biol*. 2016;14:11. DOI 10.1186/s12915-016-0231-z

Zinovyev A., Sadovsky M., Calzone L., Fouché A., Groeneveld C.S., Chervov A., Barillot E., Gorban A.N. Modeling progression of single cell populations through the cell cycle as a sequence of switches. *Front. Mol. Biosci*. 2022;8:793912. DOI 10.3389/fmolb.2021.793912

ORCID ID

S.I. Bartsev orcid.org/0000-0003-0140-4894

Acknowledgements. The study was funded by State Assignment of the Ministry of Science and Higher Education of the Russian Federation (project No. 0287-2021-0018).

I am grateful to L.Yu. Berzhanskaya for involving me in this work and to V.A. Okhonin and A.N. Shenderov for useful comments and advice in carrying out this work.

Conflict of interest. The author declares no conflict of interest.

Received July 18, 2023. Revised September 16, 2023. Accepted September 18, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):890-897
DOI 10.18699/VJGB-23-103

DyCeModel: a tool for 1D simulation for distribution of plant hormones controlling tissue patterning

D.S. Azarova[1], N.A. Omelyanchuk[1], V.V. Mironova[2], E.V. Zemlyanskaya[1, 3], V.V. Lavrekha[1, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud University, Nijmegen, the Netherlands
[3] Novosibirsk State University, Novosibirsk, Russia
vvl@bionet.nsc.ru

Abstract. To study the mechanisms of growth and development, it is necessary to analyze the dynamics of the tissue patterning regulators in time and space and to take into account their effect on the cellular dynamics within a tissue. Plant hormones are the main regulators of the cell dynamics in plant tissues; they form gradients and maxima and control molecular processes in a concentration-dependent manner. Here, we present DyCeModel, a software tool implemented in MATLAB for one-dimensional simulation of tissue with a dynamic cellular ensemble, where changes in hormone (or other active substance) concentration in the cells are described by ordinary differential equations (ODEs). We applied DyCeModel to simulate cell dynamics in plant meristems with different cellular structures and demonstrated that DyCeModel helps to identify the relationships between hormone concentration and cellular behaviors. The tool visualizes the simulation progress and presents a video obtained during the calculation. Importantly, the tool is capable of automatically adjusting the parameters by fitting the distribution of the substance concentrations predicted in the model to experimental data taken from the microscopic images. Noteworthy, DyCeModel makes it possible to build models for distinct types of plant meristems with the same ODEs, recruiting specific input characteristics for each meristem. We demonstrate the tool's efficiency by simulation of the effect of auxin and cytokinin distributions on tissue patterning in two types of *Arabidopsis thaliana* stem cell niches: the root and shoot apical meristems. The resulting models represent a promising framework for further study of the role of hormone-controlled gene regulatory networks in cell dynamics.
Key words: computer modeling; developmental trajectory; input data; genetic algorithm; phytohormones.

For citation: Azarova D.S., Omelyanchuk N.A., Mironova V.V., Zemlyanskaya E.V., Lavrekha V.V. DyCeModel: a tool for 1D simulation for distribution of plant hormones controlling tissue patterning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):890-897. DOI 10.18699/VJGB-23-103

DyCeModel: программное средство для одномерного моделирования распределения гормонов растений, контролирующих образование структуры ткани

Д.С. Азарова[1], Н.А. Омельянчук[1], В.В. Миронова[2], Е.В. Землянская[1, 3], В.В. Лавреха[1, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Университет Неймегена, Неймеген, Нидерланды
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
vvl@bionet.nsc.ru

Аннотация. Для изучения механизмов роста и развития необходимо анализировать динамику распределения регуляторов по ткани во времени и пространстве и учитывать их влияние на клеточную динамику внутри ткани. Растительные гормоны являются основными регуляторами динамики клеток в тканях растений; они образуют градиенты и максимумы и контролируют молекулярные процессы в зависимости от концентрации. Мы представляем DyCeModel, программный инструмент, реализованный в среде MATLAB для одномерного моделирования ткани с динамическим клеточным ансамблем, где изменения концентрации гормона (или другого активного вещества) в клетках описываются обыкновенными дифференциальными уравнениями. Мы применили DyCeModel для моделирования динамики клеток в меристемах растений с различной клеточной структурой и продемонстрировали, что DyCeModel помогает выявить взаимосвязь между концентрацией гормонов и поведением клеток. Инструмент визуализирует ход моделирования и предоставляет видео, полученное в ходе расчета. Важно отметить, что инструмент способен автоматически подбирать параметры, подгоняя распределение концентраций веществ, предсказанное в модели, к экспериментальным данным, полученным по изображениям с микроскопа. Примечательно, что DyCeModel позволяет строить модели для различных типов меристем растений на основе одних и тех же обыкновенных дифференциальных уравнений, используя для каждой меристемы специфические входные характеристики. Эффективность инструмента продемонстриро-

вана путем моделирования влияния распределения ауксина и цитокинина на формирование паттерна ткани в двух типах ниш стволовых клеток *Arabidopsis thaliana*: апикальных меристемах корня и побега. Полученные модели представляют собой перспективный фреймворк для дальнейшего изучения роли контролируемых гормонами генных регуляторных сетей в динамике клеток.
Ключевые слова: компьютерное моделирование; траектория развития; входные данные; генетический алгоритм; фитогормоны.

Introduction

Understanding the control of cell division and differentiation in stem cell niches is among the major issues in plant developmental biology (Hayashi et al., 2023). Although many components of the molecular regulatory networks, which underlie these processes, have been identified, complex interactions and numerous players hinder detailed study on the mechanisms of their functioning. For example, it is still largely unknown how the formation of plant hormone concentration gradients results in particular alterations in the cellular dynamics of developing tissues and organs (Rutten et al., 2022). Dissection of these issues requires application of computer modeling to predict the output in cellular dynamics and to determine whether various developmental pathways exist under certain conditions (Fisher et al., 2023).

Nowadays, developmental biology has recruited experts in mathematical modeling and computer sciences to create appropriate tools. Numerical simulations were successfully used to study the influence of phytohormone concentration distribution on the functioning of plant stem cell niches in 1D and 2D models describing cell divisions, growth, and differentiation under control of signaling molecules (Kitano et al., 2005; Nikolaev et al., 2006; Mironova et al., 2010; Muraro et al., 2013; Band et al., 2014; De Rybel et al., 2014; Lavrekha et al., 2014; Dubreuil et al., 2018; Savina et al., 2020; Hartmann et al., 2021). At the same time, these models stay within the limits of a certain meristem, and are not applicable to a wider range of plant stem cell niches. A general description of the basic set of processes related to the redistribution of hormone gradients and cellular response to this may serve as a basis for the investigation of the common and specific features of various plant meristems.

To solve this kind of problem, professional tools have started to be developed, helping researchers to create extensible computer models, which enable applying the same mathematical model equations to various plant systems (Hay Mele et al., 2015; Schölzel et al., 2021). For example, Cell Designer is a tool for simulating biochemical networks (Kitano et al., 2005) without reference to the tissue topology. A similar tool, PySB, has ample opportunity to create, extend and combine models based on genetic networks with high complexity (Lopez et al., 2013). This Python-based software is highly flexible because it provides the possibility of direct manipulation of equations. BioNetGen allows to create models both using a graphic editor and describing models manually inside the program code that simplifies reconstruction of molecular networks (Harris et al., 2016). BioNetGen has a convenient graphical representation for the solution of equations. SBMLToolbox provides the possibility to create, validate and calculate models with ODEs using SBML in MATLAB and Octave (Keating et al., 2006). DBSolve features abundances of certain molecules in a system, displaying it dynamically as a bar graph (Gizzatkulov et al., 2010). MGSmodeller is a Java application, which enables hierarchical data presentation and editing, and implements dynamic calculation tools in reconstructing molecular genetic networks and solving inverse problems (Kazantsev et al., 2008). The COPASI software is able to describe models of biological processes, such as metabolic networks, cellular signaling pathways, regulatory networks, infectious diseases and many others, simulate and analyze these models, create analysis reports and import/export models (reviewed in Bergmann et al., 2017). In COPASI, models are defined as chemical reactions between molecules. The model analyzer includes steady-state analysis, stoichiometric analysis, time history modeling using deterministic and stochastic modeling algorithms, metabolic control analysis, optimization and parameter estimation. VCell is a computing system for modeling physicochemical and electrophysiological processes in living cells (Loew, Schaff, 2001; Moraru et al., 2008). The tool allows the user to enter a description of cell physiology, biochemical reactions, and automatically or manually input mathematical equations. The resulting simulations are displayed on dynamic spatial regions of various shapes, including irregular 3D geometries derived from experimental images. VCell can also implement rule-based models, which allows the representation of species as structured objects consisting of molecules and uses reaction rules to define molecular interactions. SpringSaLaD is a software platform based on spatial stochastic modeling of biochemical systems (Michalski, Loew, 2016). SpringSaLaD models molecules as a group of connected spherical regions with excluded volume. This allows establishing a connection between molecular dynamics modeling and processes at the cellular level. SpringSaLaD is a standalone tool that supports model building, simulation, visualization, and data analysis through a graphical user interface.

The tools listed above develop models for metabolic and signal transduction pathways, and gene regulation networks. Such tools do not implement embedding of the generated mathematical models into cell ensembles to study the influence of regulatory networks on cell divisions, growth and differentiation (Kitano et al., 2005; Keating et al., 2006; Kazantsev et al., 2008; Gizzatkulov et al., 2010; Lopez et al., 2013; Harris et al., 2016).

On the other hand, there are programs that along with simulation of gene networks also consider the influence of regulatory circuits on cell growth or divisions. CompuCell3D is a tool for constructing dynamic multicellular 2D and 3D models to simulate cells that lack a cell wall (Swat et al., 2012). It is based on the lattice-based Glazier–Graner–Hogeweg (GGH) Monte Carlo multi-cell modeling, which employs an energetic approach to model growth, intercellular communication and maintenance of cell shape. Molecular processes, namely, the production and diffusion of substances, are described via ODE

solvers. The VirtualLeaf program simulates the relationship between gene expression and the biophysics of plant cell growth (Merks et al., 2011). The model is a set of cells and cell walls, through which chemical substances can move, affecting gene expression and properties of the cell wall. Cellzilla is a 2D tissue modeling platform using Cellerator, a tool describing biochemical interactions via simplified notation as reactions and converting them automatically to the corresponding differential equations by an inner computer algebra system (Shapiro et al., 2013). In Cellzilla, cells are represented by a polygonal grid of well-mixed compartments. Cell components can interact through Cellerator reactions, which describe diffusion and transport. Dynamic simulation consists of cell growth and division. Despite these advantages, modern software tools for modeling usually use manual setting of parameters, and do not support automatic parameter fitting, which may be critical for some models.

A recent trend is further improvement of computer tools, which can be used by biologists for in-depth study of developmental processes at the multicellular level. One of the current challenges is the creation of software that constructs numerical models along various plant organs utilizing uniformly described processes and provides automatic parameters setting. Here we present a tool creating one-dimensional computer models that provide embedding of signaling molecules into a dynamically developing cellular ensemble, where, based on the same set of processes, it is possible to model cellular dynamics in various plant tissues. To build realistic computer models, it is necessary to apply experimental data. The tool we have developed takes experimental data into account already at the first stage of parameter fitting, which brings the constructed models as close to reality as possible.

Materials and methods

DyCeModel overview. DyCeModel allows creating a dynamic one-dimensional cell lattice, embedding it into a mathematical model in ODE, and performing numerical analysis. It contains five script files (.m files) executed in the MATLAB software environment (Fig. 1). The substance_eq.m block incorporates an ODE system for description of synthesis, degradation, passive and active transport for the substances of interest. By default, DyCeModel provides examples of functions, which describe these processes for two substances according to Michaelis–Menten kinetics and Generalized Hill function method (Likhoshvai, Ratushny, 2007), Fick's law of diffusion and the mass action law (for describing active transport). Alternatively, users can build their own functions instead of the default ones. The parameters_fitting.m block describes the realization of a genetic algorithm to assess the similarity of the modeled substance distribution to the experimental data. The model_parameters.m block contains all model parameter default values for the ODE system and describes the model configuration of substance influxes. The grow_eq.m block describes the cell growth function, tool_1d_model.m ensures the simulation procedure. Importantly, there are two different strategies of applying DyCeModel: using the parameters_fitting.m block or omitting it. In the latter case, the user should define all parameters in the model_parameters.m file.

The input data. A pre-processed experimentally obtained microscopic image, which visualizes the distribution profile of the substance concentration within the modeled tissue, is an input for the parameters_fitting.m block. DyCeModel accepts TIFF, GIF, JPEG, PNG formats and some other graphic file formats supported by MATLAB, and it is capable of process-

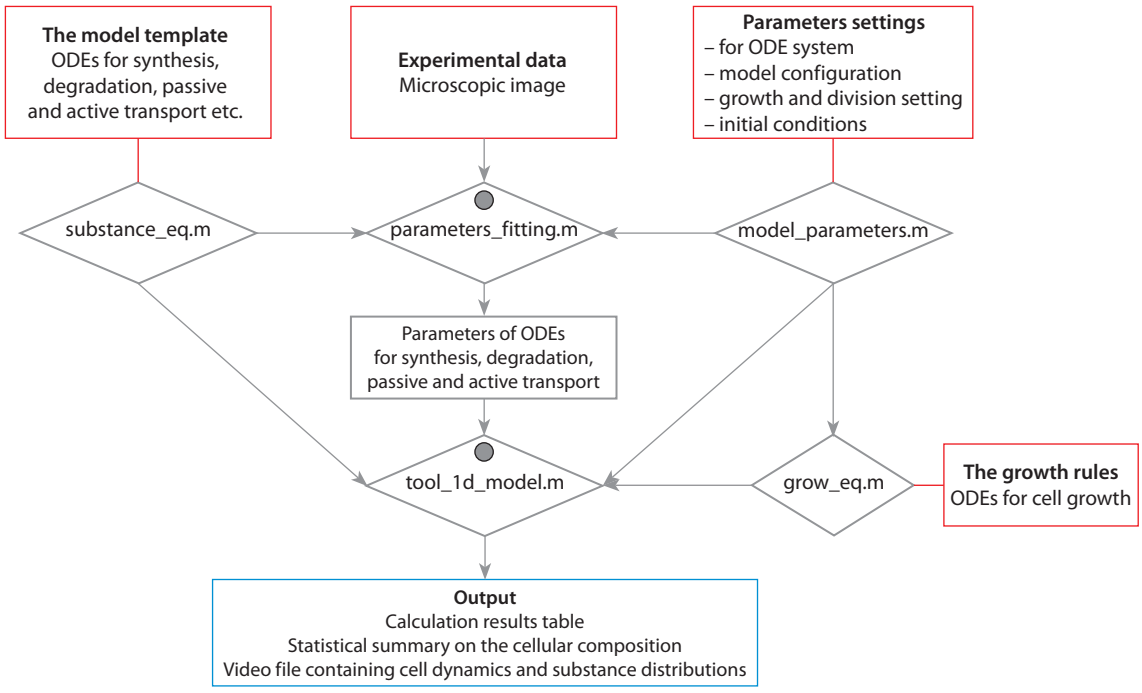


Fig. 1. DyCeModel pipeline for creating mathematical models.

The input data are marked in red. The output data are depicted in blue. Gray circles indicate the presence of visualization modules. Five script files are given in rhombuses.

ing the signal localized in the cytosol or in the nucleus. The image must be well focused. The aforementioned image preprocessing consists in excision of a rectangular area containing the modeled axis along the tissue, which should be parallel to the long side of the rectangle. This area should not contain microscope artifacts. To obtain noise-free measurements, the user can decrease the size of the rectangular area (the minimum size of the uploaded rectangular image is 1 pixel in width and 90 pixels in length). There are no strict requirements for image resolution.

The ODE system of the mathematical model is an input, which is written in the substance_eq.m file block (see Fig. 1). The default example equations can be changed according to the user's request. The model configuration is defined by the initial number of cells and position of the substance influxes in model_parameters.m file. For the model simulation, initial concentrations of all substances, as well as growth and division settings should be defined. The user also sets the number of calculation steps in order to define the time of investigation. If automatic parameter fitting is going to be omitted, the user can optionally set the parameters for the ODE system in the model_parameters.m file. All default example model parameters are consistent with the default ODE system and calculation procedure.

The parameter fitting. First, the parameters_fitting.m script quantifies the distribution of the substance concentration along the selected axis from the microscopic image. These data will be used as target distribution, which the algorithm should reproduce as precisely as possible according to the model equations and configuration. At this stage, the concentration distribution can be manually corrected if it is distorted in the microscopic image. Next, the genetic algorithm is used to find a set of model parameters, which allow reproducing the input experimental data on the distribution of the substance concentration the most accurately (Fig. 2) (Dubitzky et al., 2013).

Initially, the parameters_fitting.m script generates individuals: the sets of model parameters assigned to random values. Each individual is characterized with the fitness function value that scores the similarity of the modeled distribution of the substance concentration to the experimental data. The root-mean-square deviation (RMSD) metric is used as a fitness function. A lower fitness value corresponds to a better quality of the solution. The genetic algorithm is implemented in the following three steps.

Step 1 is "mutation", which changes a randomly selected parameter in each parameter set by the value of λ (which is also randomly selected in the interval from 0 to 1). For each individual, we calculate the model with a new parameter set. "Mutation" is fixed if it brings the solution closer to the target distribution. Step 2 is "crossover", the exchange of the parameter values between two individuals. In the first new set of parameters, a few (the number is defined randomly at each step of the algorithm) are picked from individual 1, and the rest are taken from individual 2. The second new set of parameters incorporates the values for the corresponding few parameters from individual 2, and the values for the rest of the parameters are taken from individual 1. The model is calculated with two new sets of parameters after the "crossover", and the recombination event is fixed if it brings the solution closer to the target. Step 3 supports biologically reasonable limitations on parameter values, which the user can set up manually in the "Biological limits" block of the parameters_fitting.m script (see Fig. 2). The restrictions may apply, for example, to the parity of the passive transport of different substances, the parity of active and passive transport of the same substance, the parity of the substance inflow and synthesis, etc. Taking into account reasonable biological restrictions, the algorithm "rewards" the realistic parameter values during selection, which both favors identification of the local optimum corresponding to the real processes, and speeds up the algorithm.

The fitting ends when the difference between the substance distribution calculated with the adjusted parameters and target substance distribution from the microscopic image becomes less than the threshold. The selected parameters set (the "Par" variable) is saved in a file. After executing the parameters_fitting.m script, it is recommended to inspect the selected parameters, since not all biological limitations could be taken into account during the selection. The user can view the "Par" variable and, if there are obvious inconsistencies in parameter matching, restart the parameter fitting.

Calculation of the mathematical model. When the ODE system and cell growth rules are defined, the user can load the mandatory parameters of the model with the model_parameters() function, including the initial number of cells, the initial concentrations of substances, the initial cell sizes, the maximum number of cells to be monitored, cell division parameters and cell growth settings according to the function described in the grow_eq.m file. Then the user uploads the set of parameters for the model ODE system, which are either obtained during the parameter fitting procedure or defined manually in the model_parameters.m script. After that, the model can be calculated (Fig. 3). We proceed under the assumption that cell dynamic events such as division or differentiation are discrete processes. Therefore, the calculation of ODEs is periodically interrupted to check if the conditions for cell division and differentiation specified in the tool_1d_model.m and model_parameters.m files are met. Optionally, the user decides which substances will regulate the ability to divide and the probability of cell division. All calculation results obtained during the simulation of the model are recorded in a video file, which represents the redistribution of the substance concentrations on a one-dimensional dynamic cellular ensemble.

Images used in the study. To model root apical meristem, we used publicly available images for 9-day-old *Arabidopsis thaliana* seedlings expressing *DR5::GFP* auxin sensor (Ottenschläger et al., 2003) or *TCSn::GFP* cytokinin sensor (Zürcher et al., 2013), which were obtained using a confocal fluorescence microscope (FV-1200, Olympus) (Sakamoto et al., 2019). To model shoot apical meristem, we took publicly available images for 7-day-old *A. thaliana* seedlings expressing *TCSn::GFP* cytokinin sensor obtained by a confocal microscope (Leica) (Zürcher et al., 2016). As a visualization of auxin distribution in the shoot apical meristem, we used images of auxin immunolocalization in the inflorescences of 22-day-old *A. thaliana* seedlings taken by a confocal microscope (LSM, FluoView1000, Olympus) (Banasiak et al., 2019).

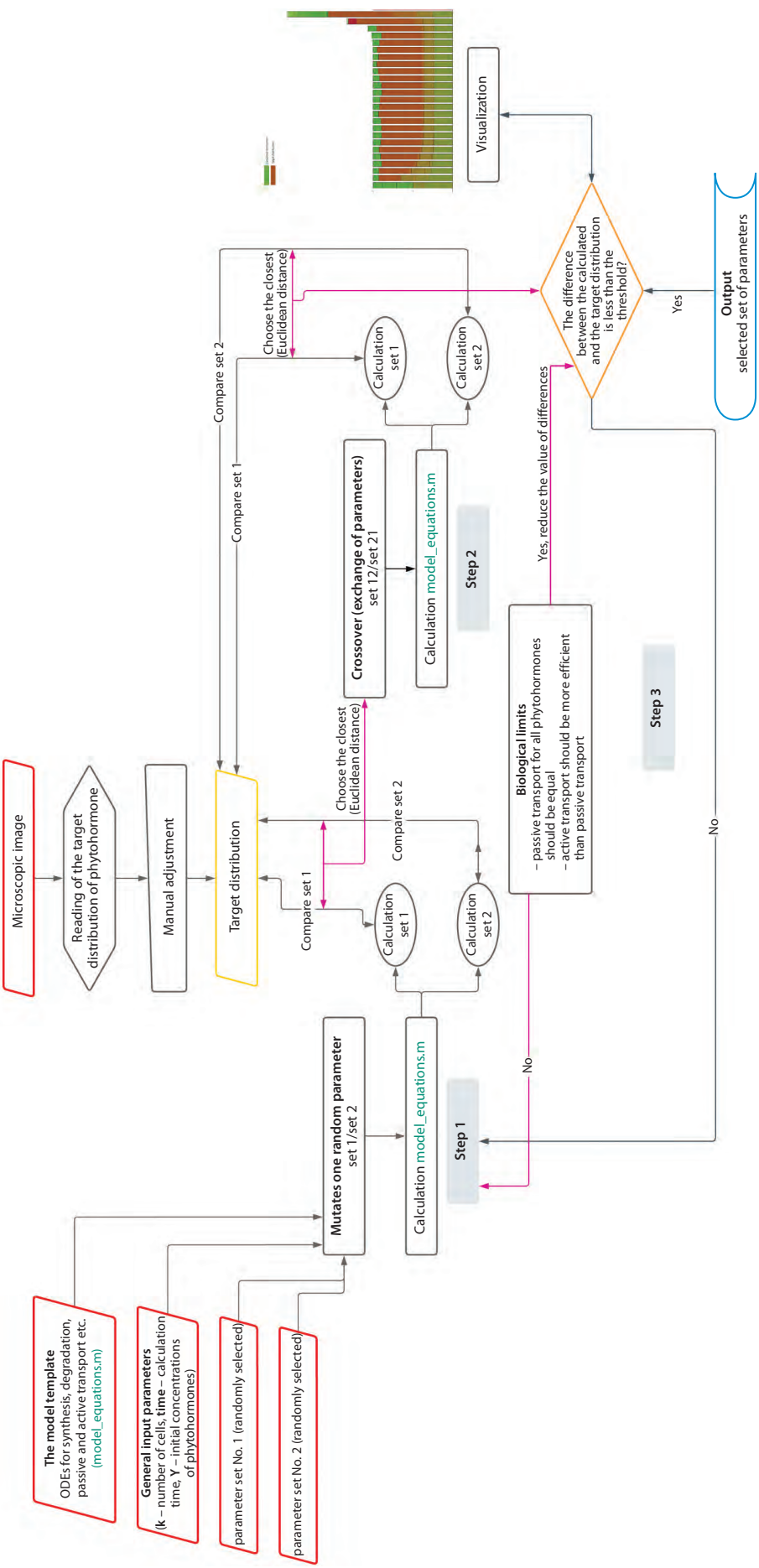


Fig. 2. The scheme of semi-automatic parameter fitting by the genetic algorithm method.
User-downloaded inputs are marked by red. Output data is depicted in blue. Data comparison blocks are marked by yellow. Black arrows indicate the processes executable in the model, pink arrows connect the parts within the comparison blocks. Orange indicates the exit block from the cyclic selection algorithm.

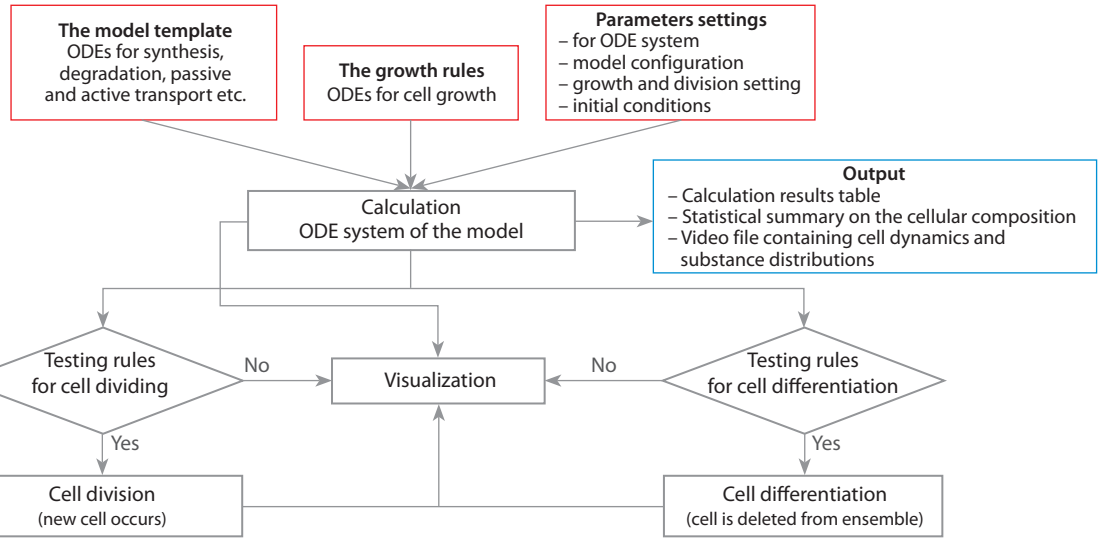


Fig. 3. The framework for calculation of the ODE system on a dynamic cell ensemble.

Input data are marked by red. Output data are depicted in blue.

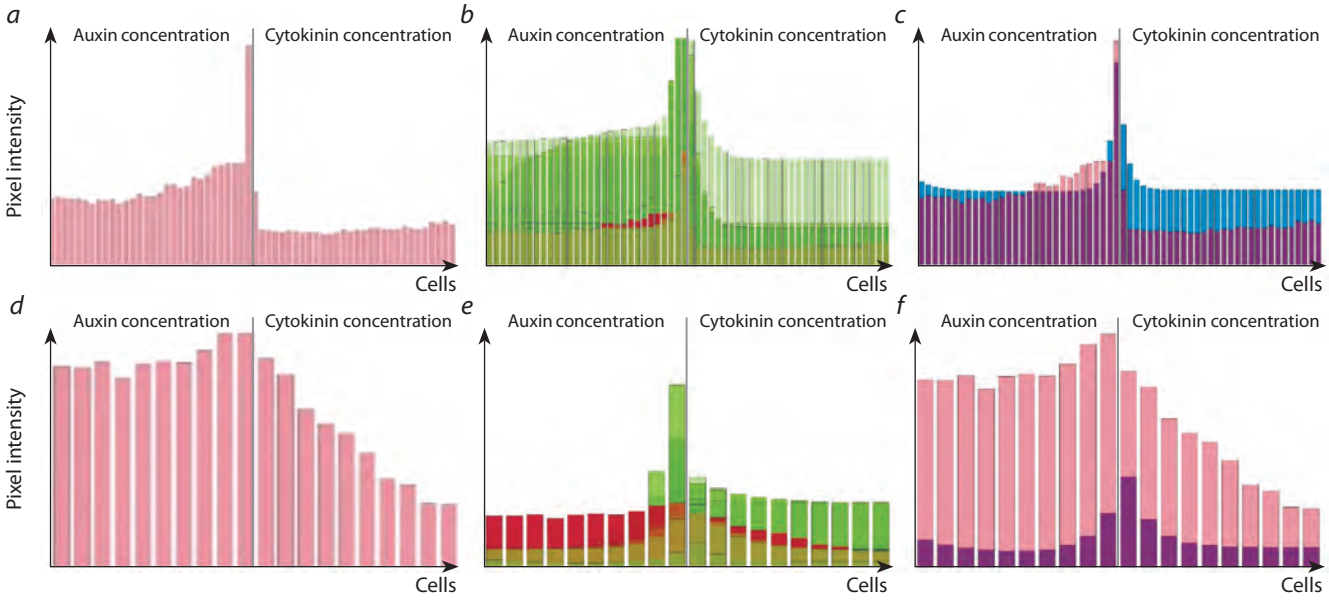


Fig. 4. DyCeModel solutions on auxin and cytokinin distribution within the root (*a–c*) and shoot (*d–f*) apical meristems along the central axis. *a*, *d*, Obtained signal intensity of hormone distributions along the allocated area; *b*, *e*, visualization of the parameter fitting process; *c*, *f*, the result of automatic parameter fitting.

Pink or burgundy indicates the signal intensity distribution obtained from the experimental data. The distributions of phytohormones during each step of parameter selection are indicated in green. Blue marks the distribution of the substances when calculating the model with the automatically selected set of parameters.

Results and discussion

A one-dimensional model of *Arabidopsis thaliana* root apical meristem built with DyCeModel

To demonstrate the performance of DyCeModel, we used it to create a 1D model of *A. thaliana* root apical meristem. Plant hormones auxin and cytokinin play major roles in regulation of maintenance of its structure (Yamoune et al., 2021). We built an ODE system based on mathematical models of auxin and cytokinin distribution published earlier (Mironova et al., 2010; Lavrekha et al., 2014). To set the parameters for the model ODE system, we used automatic parameter fitting. To define the target distribution of the hormone concentrations in the root tip, we used publicly available microscopic images described in the "Materials and methods" section. We used the following parameter value limitations during parameter fitting: approximately equal diffusion parameter values for auxin and cytokinin, prevalence of active auxin transport over the passive transport, prevalence of auxin flow into the meristem over its biosynthesis, which is typical for the root apical meristem (Overvoorde et al., 2010). Figure 4, *a–c* demonstrates auxin and cytokinin distributions in the root apical meristem gener-

ated using the DyCeModel tool. The equation describing the dependence of cell growth on auxin was built on principles similar to the Hartmann model (Hartmann et al., 2021), where the growth rate is directly proportional to auxin concentration in the cell and inversely proportional to the cell size. Cell division can occur if the cell attains minimum size required for division and possesses a certain ratio of auxin and cytokinin concentrations. The probability of cell division is 0.1, the values were obtained by analyzing images and 24-hour videos with marked division events in the meristem of *A. thaliana* (Marhava et al., 2019). Cell differentiation occurs if the cell size approaches the "maximum cell size" parameter value.

Then we built a functional model of the root apical meristem and obtained a stationary solution. Analysis of the steady-state solution of the root apical meristem model showed that it is consistent with experimental data (García-Gómez et al., 2017; Hu et al., 2021). The distribution of auxin had the shape of an inverted dome and reached a maximum in cells representing the quiescent center. The concentration of cytokinin decreased nonlinearly towards the stem cell niche and reached a minimum in the initial cells, which corresponds to experimental data. In the constructed model obtained with DyCeModel, the correct location and size of the zone of high proliferative activity were specified and remained stable for a long period of calculation, corresponding to those in the root meristem *in vivo*. Similar zones of proliferative activity were also formed in two other models of the root apical meristem (Mironova et al., 2010; Lavrekha et al., 2014).

DyCeModel enables modeling distinct types of plant meristems based on the same ODEs

We speculated that recruiting specific input characteristics for distinct meristems could enable the modeling of distinct types of plant meristems with DyCeModel based on the same ODEs. Therefore, we applied DyCeModel to build a model of *A. thaliana* shoot apical meristem using the same mathematical model equations and rules as for root apical meristem. The images used for automatic parameter fitting are described in the "Materials and methods" section. We used the following parameter value limitations during parameter fitting: approximately equal diffusion parameter values for auxin and cytokinin and a low level of auxin synthesis. In the model of the shoot apical meristem, we obtained a hormone distribution profile that qualitatively corresponds to experimental data (Heisler, Jönsson, 2006). Auxin and cytokinin concentrations decreased nonlinearly with distance from the stem cells. One-dimensional simulations of the shoot apical meristem of *A. thaliana* established a dynamic balance between dividing and differentiated cells. In this way, zones of proliferative activity were identified, and the number of cells within this zone was maintained at a certain level throughout the entire model calculation. At the same time, the identified parameters of passive transport, degradation, and growth were the same for the model of shoot meristem and root meristem, and the parameters determining cell division remained similar.

Conclusion

The DyCeModel tool constructs mathematical models of hormone distribution based on the processes of their synthesis, degradation, diffusion and active transport in a dynamically developing cellular ensemble. Such models are necessary to consider the influence of hormone distribution on cell growth and division. The developed DyCeModel tool is quite flexible, it provides embedding, addition, mixing of already existing mathematical models. Adding each model to the scripts switches on machine selection of unknown parameters, which speeds up the work with the model and makes it more stable. In addition, DyCeModel makes a statistical summary on the cellular composition that can be used for predictions about the influence of hormones on proliferative cell activity.

Using DyCeModel, we built a functional model of the root apical meristem, which was consistent with the experimental data. Next, we applied DyCeModel to build a model of the shoot apical meristem using the same mathematical model equations as for the root apical meristem model and demonstrated that the parameters of passive transport, degradation, growth, even the parameters determining cell division remain similar between root and shoot models. The resulting one-dimensional models can be further used as a framework to study the role of hormone-controlled gene networks in cell dynamics in two types of meristem.

References

Banasiak A., Biedroń M., Dolzblasz A., Berezowski M.A. Ontogenetic changes in auxin biosynthesis and distribution determine the organogenic activity of the shoot apical meristem in *pin1* mutants. *Int. J. Mol. Sci.* 2019;20(1):180. DOI 10.3390/ijms20010180

Band L.R., Wells D.M., Fozard J.A., Ghetiu T., French A.P., Pound M.P., Wilson M.H., Yu L., Li W., Hijazi H.I., Oh J., Pearce S.P., Perez-Amador M.A., Yun J., Kramer E., Alonso J.M., Godin C., Vernoux T., Hodgman T.C., Pridmore T.P., Swarup R., King J.R., Bennett M.J. Systems analysis of auxin transport in the *Arabidopsis* root apex. *Plant Cell*. 2014;26(3):862-875. DOI 10.1105/tpc.113.119495

Bergmann F.T., Hoops S., Klahn B., Kummer U., Mendes P., Pahle J., Sahle S. COPASI and its applications in biotechnology. *J. Biotechnol.* 2017;261:215-220. DOI 10.1016/j.jbiotec.2017.06.1200

De Rybel B., Adibi M., Breda A.S., Wendrich J.R., Smit M.E., Novák O., Yamaguchi N., Yoshida S., van Isterdael G., Palovaara J., Nijsse B., Boekschoten M.V., Hooiveld G., Beeckman T., Wagner D., Ljung K., Fleck C., Weijers D. Integration of growth and patterning during vascular tissue formation in *Arabidopsis*. *Science.* 2014;345(6197):1255215. DOI 10.1126/science.1255215

Dubitzky W., Wolkenhauer O., Cho K.-H., Yokota H. (Eds.) Encyclopedia of Systems Biology. New York: Springer, 2013. DOI 10.1007/978-1-4419-9863-7

Dubreuil C., Jin X., Grönlund A., Fischer U. A local auxin gradient regulates root cap self-renewal and size homeostasis. *Curr. Biol.* 2018;28(16):2581-2587.e3. DOI 10.1016/j.cub.2018.05.090

Fischer S.C., Bassel G.W., Kollmannsberger P. Tissues as networks of cells: towards generative rules of complex organ development. *J. R. Soc. Interface*. 2023;20(204):20230115. DOI 10.1098/rsif.2023.0115

García-Gómez M.L., Azpeitia E., Álvarez-Buylla E.R. A dynamic genetic-hormonal regulatory network model explains multiple cellular behaviors of the root apical meristem of *Arabidopsis thaliana*. *PLoS Comput. Biol*. 2017;13(4):e1005488. DOI 10.1371/journal.pcbi.1005488

Gizzatkulov N.M., Goryanin I.I., Metelkin E.A., Mogilevskaya E.A., Peskov K.V., Demin O.V. DBSolve Optimum: a software package for kinetic modeling which allows dynamic visualization of simulation results. *BMC Syst. Biol*. 2010;4(1):109. DOI 10.1186/1752-0509-4-109

Harris L.A., Hogg J.S., Tapia J.-J., Sekar J.A.P., Gupta S., Korsunsky I., Arora A., Barua D., Sheehan R.P., Faeder J.R. BioNetGen 2.2: ad-

vances in rule-based modeling. *Bioinformatics*. 2016;32(21):3366-3368. DOI 10.1093/bioinformatics/btw469

Hartmann F.P., Rathgeber C.B.K., Badel É., Fournier M., Moulia B. Modelling the spatial crosstalk between two biochemical signals explains wood formation dynamics and tree-ring structure. *J. Exp. Bot*. 2021;72(5):1727-1737. DOI 10.1093/jxb/eraa558

Hay Mele B., Giannino F., Vincenot C.E., Mazzoleni S., Cartení F. Cell-based models in plant developmental biology: insights into hybrid approaches. *Front. Environ. Sci*. 2015;3:73. DOI 10.3389/fenvs.2015.00073

Hayashi M., Mähönen A.P., Sakakibara H., Torii K.U., Umeda M. Plant Stem Cells: the source of plant vitality and persistent growth. *Plant Cell Physiol*. 2023;64(3):271-273. DOI 10.1093/pcp/pcad009

Heisler M.G., Jönsson H. Modeling auxin transport and plant development. *J. Plant Growth Regul*. 2006;25:302-312. DOI 10.1007/s00344-006-0066-x

Hu Y., Omary M., Hu Y., Doron O., Hoermayer L., Chen Q., Megides O., Chekli O., Ding Z., Friml J., Zhao Y., Tsarfaty I., Shani E. Cell kinetics of auxin transport and activity in *Arabidopsis* root growth and skewing. *Nat. Commun*. 2021;12(1):1657. DOI 10.1038/s41467-021-21802-3

Kazantsev F.V., Akberdin I.R., Bezmaternykhi K.D., Lashin S.A., Podkolodnaya N.N., Likhoshvai V.A. MGSmodeller – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system. In: Abstracts of the VI International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008), Novosibirsk, June 22–28. Novosibirsk: ICG, 2008;113

Keating S.M., Bornstein B.J., Finney A., Hucka M. SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics*. 2006;22(10):1275-1277. DOI 10.1093/bioinformatics/btl111

Kitano H., Funahashi A., Matsuoka Y., Oda K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol*. 2005;23(8):961-966. DOI 10.1038/nbt1111

Lavrekha V.V., Omelyanchuk N.A., Mironova V.V. Mathematical model of phytohormone regulation of root meristematic zone formation. *Vavilov J. Genet. Breed*. 2014;18(4/2):963-972 (in Russian)

Likhoshvai V., Ratushny A. Generalized hill function method for modeling molecular processes. *J. Bioinform. Comput. Biol*. 2007;5(2B):521-531. DOI 10.1142/s0219720007002837

Loew L.M., Schaff J.C. The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol*. 2001;19(10):401-406. DOI 10.1016/S0167-7799(01)01740-1

Lopez C.F., Muhlich J.L., Bachman J.A., Sorger P.K. Programming biological models in Python using PySB. *Mol. Syst. Biol*. 2013;9(1):646. DOI 10.1038/msb.2013.1

Marhava P., Hoermayer L., Yoshida S., Marhavý P., Benková E., Friml J. Re-activation of stem cell pathways for pattern restoration in plant wound healing. *Cell*. 2019;177(4):957-969.e13. DOI 10.1016/j.cell.2019.04.015

Merks R.M.H., Guravage M., Inzé D., Beemster G.T.S. VirtualLeaf: an open-source framework for cell-based modeling of plant tissue growth and development. *Plant Physiol*. 2011;155(2):656-666. DOI 10.1104/pp.110.167619

Michalski P.J., Loew L.M. SpringSaLaD: a spatial, particle-based biochemical simulation platform with excluded volume. *Biophys. J*. 2016;110(3):523-529. DOI 10.1016/j.bpj.2015.12.026

Mironova V.V., Omelyanchuk N.A., Yosiphon G., Fadeev S.I., Kolchanov N.A., Mjolsness E., Likhoshvai V.A. A plausible mechanism for auxin patterning along the developing root. *BMC Syst. Biol*. 2010;4(1):98. DOI 10.1186/1752-0509-4-98

Moraru I.I., Schaff J.C., Slepchenko B.M., Blinov M.L., Morgan F., Lakshminarayana A., Gao F., Li Y., Loew L.M. Virtual Cell modelling and simulation software environment. *IET Syst. Biol*. 2008;2(5):352-362. DOI 10.1049/iet-syb:20080102

Muraro D., Byrne H., King J., Bennett M. The role of auxin and cytokinin signalling in specifying the root architecture of *Arabidopsis thaliana*. *J. Theor. Biol*. 2013;317:71-86. DOI 10.1016/j.jtbi.2012.08.032

Nikolaev S.V., Kolchanov N.A., Fadeev S.I., Kogai V.V., Mjolsness E. Investigation of a one-dimensional model of the regulation of the size of the renewal zone in biological tissue, taking into account cell division. *Computational Technologies*. 2006;11(2):67-81. (in Russian)

Ottenschläger I., Wolff P., Wolverton C., Bhalerao R.P., Sandberg G., Ishikawa H., Evans M., Palme K. Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl. Acad. Sci. USA*. 2003;100(5):2987-2991. DOI 10.1073/pnas.0437936100

Overvoorde P., Fukaki H., Beeckman T. Auxin control of root development. *Cold Spring Harb. Perspect. Biol*. 2010;2(6):a001537. DOI 10.1101/cshperspect.a001537

Rutten J., van den Berg T., Tusscher K.T. Modeling auxin signaling in roots: auxin computations. *Cold Spring Harb. Perspect. Biol*. 2022;14(2):a040089. DOI 10.1101/cshperspect.a040089

Sakamoto T., Sotta N., Suzuki T., Fujiwara T., Matsunaga S. The 26S proteasome is required for the maintenance of root apical meristem by modulating auxin and cytokinin responses under high-boron stress. *Front. Plant Sci*. 2019;10:590. DOI 10.3389/fpls.2019.00590

Savina M.S., Pasternak T., Omelyanchuk N.A., Novikova D.D., Palme K., Mironova V.V., Lavrekha V.V. Cell dynamics in WOX5-overexpressing root tips: the impact of local auxin biosynthesis. *Front. Plant Sci*. 2020;11:560169. DOI 10.3389/fpls.2020.560169

Schölzel C., Blesius V., Ernst G., Goesmann A., Dominik A. Countering reproducibility issues in mathematical models with software engineering techniques: a case study using a one-dimensional mathematical model of the atrioventricular node. *PLoS One*. 2021;16(7):e0254749. DOI 10.1371/journal.pone.0254749

Shapiro B.E., Meyerowitz E.M., Mjolsness E. Using Cellzilla for plant growth simulations at the cellular level. *Front. Plant Sci*. 2013;4:408. DOI 10.3389/fpls.2013.00408

Swat M.H., Thomas G.L., Belmonte J.M., Shirinifard A., Hmeljak D., Glazier J.A. Multi-scale modeling of tissues using CompuCell3D. *Methods Cell Biol*. 2012;110:325-366. DOI 10.1016/B978-0-12-388403-9.00013-8

Yamoune A., Cuyacot A.R., Zdarska M., Hejatko J. Hormonal orchestration of root apical meristem formation and maintenance in Arabidopsis. *J. Exp. Bot*. 2021;72(19):6768-6788. DOI 10.1093/jxb/erab360

Zürcher E., Tavor-Deslex D., Lituiev D., Enkerli K., Tarr P.T., Müller B. A robust and sensitive synthetic sensor to monitor the transcriptional output of the cytokinin signaling network in planta. *Plant Physiol*. 2013;161(3):1066-1075. DOI 10.1104/pp.112.211763

Zürcher E., Liu J., di Donato M., Geisler M., Müller B. Plant development regulated by cytokinin sinks. *Science*. 2016;353(6303):1027-1030. DOI 10.1126/science.aaf7254

ORCID ID

D.S. Azarova orcid.org/0009-0006-2030-6842
V.V. Mironova orcid.org/0000-0003-3438-0147
E.V. Zemlyanskaya orcid.org/0009-0005-7316-7690
V.V. Lavrekha orcid.org/0000-0001-8813-8941

Acknowledgements. The model development was supported by the budget project FWNR-2022-0020. All computational experiments were supported by the Russian Science Foundation, grant No. 20-14-00140.

Conflict of interest. The authors declare no conflict of interest.

Received August 16, 2023. Revised October 2, 2023. Accepted October 5, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):898-905
DOI 10.18699/VJGB-23-104

Original Russian text https://vavilovj-icg.ru/

Laboratory information systems for research management in biology

A.M. Mukhin[1, 2, 3] ✉, F.V. Kazantsev[1, 2, 3], S.A. Lashin[1, 2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ mukhin@bionet.nsc.ru

Abstract. Modern investigations in biology often require the efforts of one or more groups of researchers. Often these are groups of specialists from various scientific fields who generate and share data of different formats and sizes. Without modern approaches to work automation and data versioning (where data from different collaborators are stored at different points in time), teamwork quickly devolves into unmanageable confusion. In this review, we present a number of information systems designed to solve these problems. Their application to the organization of scientific activity helps to manage the flow of actions and data, allowing all participants to work with relevant information and solving the issue of reproducibility of both experimental and computational results. The article describes methods for organizing data flows within a team, principles for organizing metadata and ontologies. The information systems Trello, Git, Redmine, SEEK, OpenBIS and Galaxy are considered. Their functionality and scope of use are described. Before using any tools, it is important to understand the purpose of implementation, to define the set of tasks they should solve, and, based on this, to formulate requirements and finally to monitor the application of recommendations in the field. The tasks of creating a framework of ontologies, metadata, data warehousing schemas and software systems are key for a team that has decided to undertake work to automate data circulation. It is not always possible to implement such systems in their entirety, but one should still strive to do so through a step-by-step introduction of principles for organizing data and tasks with the mastery of individual software tools. It is worth noting that Trello, Git, and Redmine are easier to use, customize, and support for small research groups. At the same time, SEEK, OpenBIS, and Galaxy are more specific and their use is advisable if the capabilities of simple systems are no longer sufficient.
Key words: management; LIMS; ELN; FAIR; version control systems; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

For citation: Mukhin A.M., Kazantsev F.V., Lashin S.A. Laboratory information systems for research management in biology. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):898-905. DOI 10.18699/VJGB-23-104

Лабораторные информационные системы для управления исследовательскими работами в биологии

А.М. Мухин[1, 2, 3] ✉, Ф.В. Казанцев[1, 2, 3], С.А. Лашин[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ mukhin@bionet.nsc.ru

Аннотация. Современная исследовательская работа в биологии нередко требует усилий одной или нескольких групп исследователей. Часто это группы специалистов из смежных областей, которые генерируют и обмениваются данными разных форматов и размеров. Без применения современных подходов автоматизации работы и версионирования данных (когда данные от разных сотрудников сохраняются в разные моменты времени) коллективная работа быстро переходит в неуправляемый хаос. В настоящем обзоре приведен ряд информационных систем, предназначенных для решения озвученных задач. Их применение для организации научной деятельности позволяет управлять потоком действий и данных, добиваясь работы всех участников с актуальной информацией, и решением вопроса воспроизводимости как экспериментальных, так и вычислительных результатов. Описаны методики по организации потоков данных в рамках работы коллектива, принципы по организации метаданных и онтологий. Рассмотрены информационные системы Trello, Git, Redmine, SEEK, OpenBIS и Galaxy. Описана их функциональность и сфера использования. Выбирая те или иные инструменты, важно понимать цель внедрения, определить набор задач, которые они должны решать, и исходя из этого формулировать требования и отслеживать применение рекомендаций на местах.

Задачи по созданию структуры онтологий, метаданных, схем хранения данных и программных систем являются ключевыми для коллектива, который решился на проведение работ по автоматизации оборота данных. Не всегда возможно внедрить такие системы целиком, но все же следует стремиться к этому через поэтапное внедрение принципов по организации данных и задач с освоением отдельных программных инструментов. Следует отметить, что системы Trello, Git и Redmine проще в использовании, настройке и поддержке для малых исследовательских групп. В то же время SEEK, OpenBIS и Galaxy более специфичные, их применение целесообразно в случае, если возможностей простых систем уже недостаточно.
Ключевые слова: управление; LIMS; ELN; FAIR; системы контроля версий; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

Introduction

Modern research work in biology often requires the efforts of one or more groups of researchers. Often, these are groups of specialists from related fields who generate and exchange data of different formats and sizes. To automate and computerize this work, various tools are used to catalog, log the progress of experiments, and record results: paper notebooks and laboratory journals, spreadsheet programs, and report writing in various text editors. Without the use of modern approaches of work automation and data versioning, the team quickly succumbs to "uncontrollable chaos". A critical point in the organization of interaction in the team is the complexity of the procedure of knowledge transfer from one team member to another, as such knowledge is not formalized and often contains notes understandable only to the author. All this leads to delays in the next stages of research or in the design of publications. Sometimes employees forget to record new facts and notes, or do not keep any records of intermediate stages of work at all. This leads to irretrievable loss of knowledge and waste of resources for repeated experiments and observations.

When collecting primary data, researchers may also make errors in processing values or assigning them to a particular category. For example, transcriptome data may be erroneously attributed to a different organism from the one from which it was obtained; data may not be recorded in a uniform manner, using values of different types (integer, floating point number, string, date, etc.). Also, when working with Excel, strings may be mistakenly converted to floating point numbers, which can be critical to the interpretation of the study results (Zeeberg et al., 2004), so implicit data conversions should be avoided. In (Roche et al., 2015), Bioresource Collections (BRCs) in Ecology and Evolution were analyzed. It was found that 56 % of these BRCs were incomplete, i. e. there were blank values in the tabular data, and 64 % were collected in such a way that it was not possible to reuse the stored data due to errors in recording values.

Therefore, every team faces the task of properly formalizing the processes of data management and knowledge sharing between employees. In the following article, we will look at specific data organization methodologies and information systems and software tools that implement them, which are used by scientific organizations to distribute tasks and automate the flow of work data.

Data and process organization methodologies

There are several ways to address the challenge of organizing scientific data flows, but all require the research team to create systems of arrangements for managing, processing, and communicating scientific information. Automation systems with managed access help preserve knowledge, regulations and other "substances" of laboratory work, and do not require constant coordination. The following issues arise at the outset of these activities: (1) use of existing data design standards developed by the professional community; (2) formalization or creation of a common "working language" within the team; (3) deployment, implementation and maintenance of the information system and setting up access rights for user groups.

Transition to the use of existing standards and formats for data representation or the creation of one's own formats with comprehensive documentation sufficient for unambiguous interpretation of values allows to overcome the problem of knowledge transfer between employees inside and outside the team. Supporting documentation will be used to automate work with the information system, for example, to build modules for generating summary diagrams and reports. Formal schemes for describing the results of scientific activity have recently become widely used for fast information retrieval and interpretation of these files not only by machines but also by people. Examples include mathematical models in SBML (Hucka et al., 2019), SBGN (Novère et al., 2009) formats supported by the CO.MBINE community (Schreiber et al., 2015). We also note the MIRIAM approach for describing holistic biochemical systems (Novère et al., 2005) and the MIAME format (Brazma et al., 2001) for describing sequencing results on microarrays or RNA sequences.

When data representation standards are defined, the stage comes to formalize or create a common working language and exchange protocols within the team to streamline the transfer of subject matter knowledge. If we leave the "as it is convenient/as it was before" approach to data presentation, the issue of ambiguous or missing knowledge in the database will not be solved, which will lead to additional resource costs for correcting data at later stages of work. Ontology tools (Guizzardi, 2020) can help in solving the problem of formalization and creating a common working language. Ontologies are a broader class of knowledge organization systems for describing results in comparison to the aforementioned formal schemas. In ontology systems, it is possible to establish "concepts" and "relations" between concepts, rather than

strictly follow a ready-made schema proposed by someone earlier. Ontologies are created in order to describe meaningful information and to unambiguously interpret a system of concepts and processes within and outside the team. Teams use both simple methods to describe ontologies, such as first-order logic language, and more complex tree structures, such as OntoUML (Guizzardi et al., 2018) or RDF schemas (Gutierrez et al., 2007). Also, mathematical category theory (Kuś, Skowron, 2019) is gaining popularity for composing ontological relationships of a subject domain, which is designed to connect different areas of mathematics and subject domains with each other. A graphical language of "ontology logs" (English "ologs", essentially descriptions of a subject area in the form of graphs, where nodes describe objects with certain properties and edges describe functions of transforming one object into another) has also been implemented using the foundations of this theory (Spivak, Kent, 2012). Currently, the tools and language of category theory are not widely used in scientific publications and systems, but there are works on the use of this language in neurobiology (Brown, Porter, 2008) and on the mathematical description of an evolving model of memory (Ehresmann, Vanbremeersch, 2007).

One way to formalize the stages of laboratory work is to create metadata – information describing the data themselves (Roche et al., 2015). The format of their description is rather free. Metadata can be described/represented in the form of a structured file (XML or JSON) or database tables of both relational (Postgrespro.ru) and document-oriented structure (MongoDB.com). The description can contain any information, such as what the columns in the tables mean, what units of measurement are used, from which organism the materials were collected, how these results were obtained. Metadata can be used in conjunction with ontology systems and formal schemes for representing scientific results for quick retrieval of relevant information and unambiguous interpretation of results.

The FAIR research community (Wilkinson et al., 2016) proposed their set of principles for describing data and metadata in the task of storing and transferring information both between teams of researchers and between different data analysis programs. They formulated the following four principles that a laboratory information system should possess:

1. Findable – (meta)data are unique and uniquely identifiable. The system should have a basic mechanism for reading a detailed description, and should be able to search for these data by key fields.
2. Accessible – the data are readable by both humans and computers for further work. This is achieved using standard formats and protocols.
3. Interoperable – (meta)data are described in a machine-readable form, in a usable format and are annotated using ontologies.
4. Reusable – (meta)data are sufficiently well described so that they can be shared with other people and systems for further analysis. This item is a logical consequence of the above items.

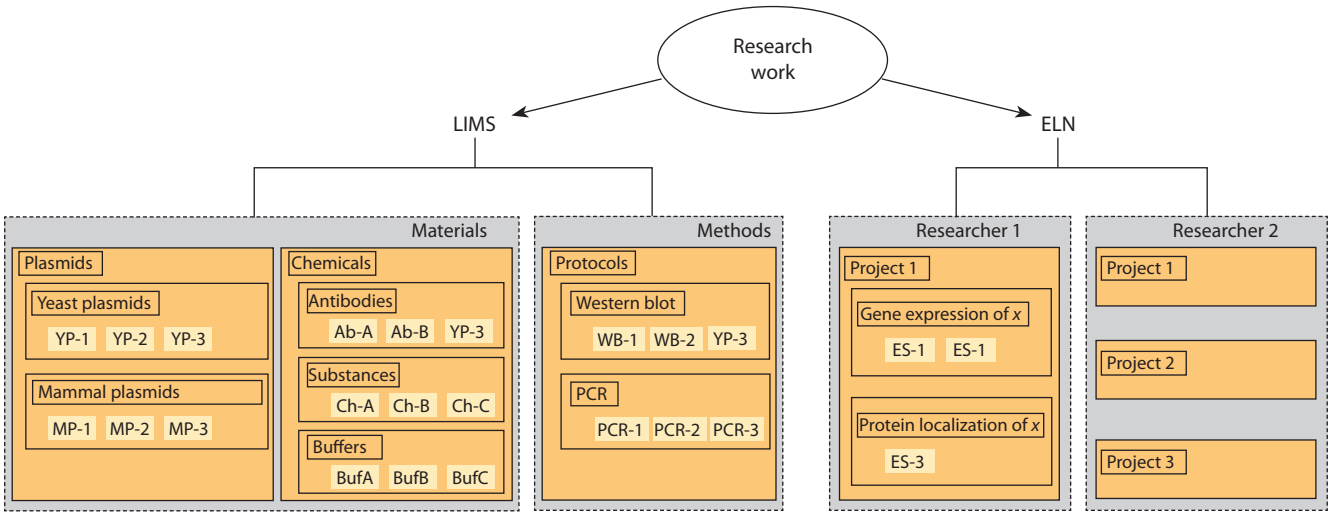
Next, let's look at software tools that are worth considering for solving data management challenges on the path to research automation.

Software tools

Two concepts, LIMS and ELN (Barillari et al., 2016), which are implemented in software packages for research control tasks, are shown in the Figure.

LIMS (Laboratory Information Management System) is a laboratory information management system. Its tasks include the management and control of laboratory materials and methods. With the help of this system, researchers can carry out document management with administration, and companies, create schedules for the use of instruments, record reagents, research objects, etc.

ELN (Electronic Laboratory Notebook) is an electronic laboratory journal. The tasks of such systems include management of projects, experiments, users, research groups, as well as logging (journaling) and control of experiments. In essence, these systems replace the functions of paper



Description of the data structure stored in LIMS and ELN systems.

notebooks for entering and transmitting notes as experiments progress.

Trello

Trello (https://trello.com/) is a conditionally free web-based workflow and communication service. In this system, users set up a virtual whiteboard on which "cards" with "tasks" are placed. The board itself is divided into sections, between which the cards are moved, showing the movement through the work stages. Most often, the board sections are marked with the statuses of work execution, for example, "tasks in queue", "in progress", "waiting for feedback", "task completed". It is possible to independently create sections according to your own scenario, which best reflects the workflow of the team. In this way, employees and managers can: (1) monitor the progress of work in real time; (2) change the statuses of tasks, add comments to tasks; (3) link tasks to each other; (4) react at early stages in cases of suspended work.

The disadvantages of Trello include the inability to modify the functionality of the system with its own modules and limited functionality in the free version. The analogs include Yandex.Tracker (https://cloud.yandex.ru/services/tracker), GutHub Projects (https://docs.github.com/en/issues/planning-and-tracking-with-projects/learning-about-projects/quickstart-for-projects) and Kanboard (https://kanboard.org/). The proposed tools are focused on the implementation of ELN requirements, but users can adapt them to LIMS tasks. They are aimed at managing team processes – other tools should be used to organize the storage and movement of the data themselves.

GitHub

When a team works together on program codes, documents and reports, there is an important task of change control. Mail clients and person-to-person networking do not cope well with this task, as the users themselves need to control the relevance of the versions of these documents. Also, the task of versioning data and text is not solved due to the lack of a system for centralizing the storage of files and fixing their changes. It is these tasks that can be solved by using the Git program (Chacon, Straub, 2014).

The Git program creates repository files in a local folder, allowing you to navigate between changes to the files. This system is most often used by programmers to work on a project simultaneously, comparing and merging code changes from different developers. Open source projects are most often stored publicly on the GitHub project servers (https://github.com). Some research groups use the Git version control system to produce articles and dissertations. For example, it was used to write a mathematical book on homotopy type theory (The Univalent Foundations Program, 2013). About 20 people worked on the book, and the cloud storage service Dropbox could not cope with the task of synchronizing the text. As a result, the team produced a 600-page book in less than six months (https://math.andrej.com/2013/06/20/the-hott-book/).

GitHub itself cannot be installed on a local computer, but there are similar solutions that can be installed on a local system, such as GitLab (https://gitlab.com), Gogs (https://gogs.io), Gitea (https://gitea.com), and GitWeb (https://git-scm.com/docs/gitweb). Within these systems, it is possible to solve ELN and LIMS tasks, but users will have to understand Git in detail.

Redmine

Redmine (https://redmine.org/) is widely used as a project control and task assignment system. Most often, the main project manager (administrator, head of laboratory, etc.) creates a set of tasks and assigns responsible executors. The executors change the status of task readiness as they complete the task. The system automatically monitors the status of project tasks and builds summary diagrams that show the time discrepancy between the plan and the actual execution. Also, the main functions of this system include:

- Role creation and restriction – the administrator can create several additional user roles and set rules for them to work in the system (reading and/or writing "tasks", wiki pages and so on).
- Flexible error control system – the function is widely used in software development, when testers or users add a "task" of the "error" type to the system to notify developers.
- Calendar and Gantt Chart. They are used to keep track of task due dates.
- Adding project news with notification of participants.
- Adding documents and files to the system.
- Notifying users by e-mail or RSS feed.
- Formalization of knowledge for each project in the format of Wikipedia – an electronic encyclopedia/reference book in the form of Internet pages.
- Forum system for each project – the ability to publicly discuss in one place the solution of problems; the ability to quickly run your eyes over the chains of messages on the topic.
- Time accounting of work on tasks and the project as a whole.
- Creation of user forms and fields for additional description of "tasks", "projects", "users" and other entities within this system.

The system can be deployed in a local information environment (up to a personal computer). It is possible to add new functionality through the implementation of submodules (plug-ins). The disadvantages of Redmine include the absence of a task board like Trello, which is clear and easy to use, as well as the limited functionality of the standard version. Therefore, for full-fledged work, it is necessary to install third-party submodules.

Many teams in the IT sector have built their workflows on the basis of the Redmine software system. In 2019, the ENVRI-FAIR project (Petzold et al., 2019) was launched to connect resources and data between the European Environmental Research Infrastructure (ENVRI) cluster and the European Open Science Computing Cloud (EOSC) using Redmine (this information was obtained from the technical documentation of this project). Based on Redmine, it is possible to realize the solution of both ELN and LIMS tasks.

SEEK system

The SEEK system (Wolstencroft et al., 2015) is designed to manage, disseminate, and explore mathematical models and associated systems biology data. SEEK organizes research project information including experimental data and bioinformatics results within a structure of three entities: Investigation, Stage, Assay (ISA) (Rocca-Serra et al., 2010). Investigation reveals the essence of a particular project (who is doing the work, which institute, the time of the study). Stage describes a specific stage in the course of the study (excretion of DNA or protein from the tissue of the organism under study, mapping of RNA reads to a reference genome, etc.). Assay is the unit of the result of the work performed. Also, in the system, it is possible to establish associative connection between samples.

The advantage of this system is the linking of data with each other within the above structure with the description of the research team, as well as the reformatting of metadata into an RDF knowledge graph (Gutierrez et al., 2007) using Virtuoso server (Software, 2022). Metadata are described mainly in tabular form (ISA-Tab), there is also a possibility to use JSON schema. For manual annotation of data, SEEK developers suggest using FightField software. Search of RDF graph data using SPARQL query language is flexible in use – in comparison to SQL, where, in addition to writing data selection rules, the user is required to manually describe the list of tables and the way they are joined. Another problem with SQL is that the user has to optimize their queries to perform searches quickly.

The main focus of SEEK is the storage and transfer of mathematical models of biological processes, the resource also allows working with SBML models and opening them in JWS Online (Olivier, Snoep, 2004) and in COPASI (Hoops et al., 2006). This system mainly implements the ELN requirements for bioinformatics projects and LIMS is not implemented in it.

OpenBIS system

As part of the laboratory work, researchers are tasked with creating protocols of experiments, following these protocols and fixing the results of work, fixing events, etc. There is a need to align the results of a series of works within a single project, for example, linking experiments to different organisms, their phenotypes, genotypes, developmental environment and other data. OpenBIS (Bauch et al., 2011) provides functionality to store and align metadata under detailed descriptions of experiments, their results, parameters, etc. The OpenBIS system consists of three modules: application server, data server and metadata database.

- The application server is the access point for users. The module provides access to the program complex through a graphical user interface, as well as via HTTP protocol (OpenBIS provides libraries in the Python, Java and Matlab programming languages for interaction over the network). To add new functions (e. g., mass spectrometry data storage), OpenBIS provides a system of modules, each of which must be implemented in the Jython programming language. This module divides authorization among users (read data, read/write data).
- The data server performs the work of organizing primary data storage on disk drives.
- The metadata database is a PostgreSQL database management system (DBMS). This module links data in projects, stores metadata, points to data from the data server, provides data search tasks.
- The ability to link to data on external resources (BigDataLink module). Metadata are stored in the metadata base, while the original information is not stored on the data server, but remains on external resources. This function is used when working with large files.
- Extension of functionality using libraries in Java, Python, JavaScript, Matlab for interaction with the OpenBIS system (data retrieval/downloading, metadata search). These libraries use hardware interface REST API of OpenBIS service, so it is possible to realize modules for interaction with the system in other programming languages. It can be used for realization of automated calculations with attraction of stored data from the OpenBIS system.
- The structure of data storage is hierarchical and organized as follows: space, project, experiment/collection, Object/Sample, Data Set.
- To link objects and data with each other, there is a method to establish ancestor–descendant relationships, i. e. the system can create a graph of objects and data.
- Import/export of data in tabular form.
- Realization of additional functionality of the system itself by means of a system of modules.
- The system performs audit of each change in its databases.
- Semantic annotation of data – description of results in a convenient and interpretable format. An RDF schema (Gutierrez et al., 2007) is used to describe the semantics.
- Integration with the SEEK system.

This system has proven itself for primary storage of biological information obtained during experiments. In (Friedrich et al., 2015), a system was implemented to add and record experimental data on different tissues of organisms when different drugs were administered. At the first level of the storage system, the object of study is described (e. g., a particular mouse in the laboratory that has been injected with a particular drug). The second level describes the particular biological tissue that was extracted from the subject. The third level describes the sequences (nucleotide or protein sequences) extracted from the object tissue under study. The system is based on LIMS and ELN requirements and is an exemplary implementation of them.

Galaxy

The systems described above are mainly systems for controlling laboratory data, but the challenges for bioinformatics laboratories are exactly the same: control of data flow, reproducibility of calculations, access to data and their storage in the server. The Galaxy system (Galaxy Community, 2022) has been implemented to solve such problems. Galaxy consists of the following modules: (1) a server with a software and GUI interface and (2) workflows, which run analytic pipelines at

the request of users. Users can either independently run the programs installed in the server and store their data (sequences, annotations, protein lists, etc.) there.

Computational pipelines can be implemented in the form of a graph, where the vertices denote programs with configured parameters, and they are connected by edges that denote the direction of data movement from the output of one program to the input of another. These processes can also run programs on a remote server or cluster, and exchange files through a common file system. Reproducibility of computational programs is achieved using the Conda environment system (Yan Y., Yan J., 2018), where an independent environment (a set of libraries, programs, and modules in Python\R of strictly defined versions) is created for each program. The lightweight virtualization system Docker (Rad et al., 2017) can also be used, where the program runs in a "virtual" and "lightweight" operating system of the Linux family. Galaxy is a FAIR-like system (Hiltemann et al., 2023). In essence, Galaxy implements an ELN requirements system but in the domain of bioinformatics pipelines, i. e., it is not a full ELN. LIMS is not fully implemented, there is only multi-user input and a limitation on the storage of computational results.

Conclusion

In this paper, a limited number of information solutions in the field of organization of project activities of laboratories working in the field of biology were considered. The Table describes brief characteristics of the reviewed systems. Such solutions as OpenBIS, SEEK and Galaxy were created specifically for scientific work, while Trello and Redmine are project management systems of more general categories, although they can be used in the work of scientific groups. The Git software suite can be considered by large teams as a tool for sharing and versioning program code, data, article texts, monographs, and other scientific texts. It should be noted that Git is not intended for storing binary files (in particular, files in DOCX, PDF, etc.), as it only considers changes to text files. Markdown and LaTeX are more appropriate formats for this use of Git.

Before implementing these or those tools, it is important to understand the goals of their implementation; based on the goals, it is important to formulate requirements, define a set of tasks to be solved by the system, and monitor the application of recommendations by specific implementers. Taking into account the complexity of the above processes, it can be

Comparison of software solutions

System name	Main field of work	Hierarchy levels	Metadata usage	LIMS	ELN	Development tools used	Deployments
Trello	Organizing tasks as notes on the board (Kanban style)	Project Stages Task		–/+	+	Cannot be installed in a local environment	Does not need to be deployed, other tools (e. g. Kanboard) are required for local deployment
Git	Versioning text data	Free	Change files and commit tree	–/+	+/–	The git application, GitHub cannot be installed locally	Effective operation requires git skills and either using a third-party GitHub service or deploying a local server (GitLab, Gitea)
Redmine	Organization of work on projects (used in IT)	Project Tasks	PostgreSQL server You can add custom fields for description	+/–	+	Ruby, PostgreSQL	System and database deployment required
OpenBIS	Laboratory management (LIMS) and Project management (ELN)	Projects Experiments Samples Dataset	PostgreSQL server Custom fields	+	+	Java, PostgreSQL	System and database deployment required
SEEK	Systems biology data and model management (ELN)	ISA standard: Investigation Stage Assay	RDF schemas Custom fields at the Assay level	–	+	Ruby, MySQL, Virtuoso	System and database deployment required
Galaxy	Reproducibility of computational experiments/protocols	Absent, there is a relationship between the data	PostgreSQL metadata database	–	+	Python, PostgreSQL	System and database deployment is required, cluster setup is also required

recommended to start with implementation from open formats and standards for presentation and transmission of biological data proposed and developed by the scientific community. The use of general-purpose workflow systems in the laboratory will allow obtaining operational experience, which, in turn, will help to determine the data formats, work protocols, and software products required for the laboratory, and, based on this, to make a decision on scaling the automation of work with data, including the creation of ontology structures, metadata, storage schemes, and scenarios for the operation of software systems.

References

Barillari C., Ottoz D.S.M., Fuentes-Serna J.M., Ramakrishnan C., Rinn B., Rudolf F. openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics.* 2016;32(4):638-640. DOI 10.1093/bioinformatics/btv606

Bauch A., Adamczyk I., Buczek P., Elmer F.J., Enimanev K., Glyzewski P., Kohler M., Pylak T., Quandt A., Ramakrishnan C., Beisel C., Malmström L., Aebersold R., Rinn B. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*. 2011;12:468. DOI 10.1186/1471-2105-12-468

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., Gaasterland T., Glenisson P., Holstege F.C., Kim I.F., Markowitz V., Matese J.C., Parkinson H., Robinson A., Sarkans U., Schulze-Kremer S., Stewart J., Taylor R., Vilo J., Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. 2001;29(4):365-371. DOI 10.1038/ng1201-365.

Brown R., Porter T. Category Theory and Higher Dimensional Algebra: potential descriptive tools in neuroscience. *arXiv*. 2003. DOI 10.48550/arXiv.math/0306223

Chacon S., Straub B. Pro Git. Kaliforniya: Apress Berkli, 2014. DOI 10.1007/978-1-4842-0076-6

Ehresmann A., Vanbremeersch J. Memory Evolutive Systems: Hierarchy, Emergence, Cognition. Elsevier Science, 2007.

Friedrich A., Kenar E., Kohlbacher O., Nahnsen S. Intuitive web-based experimental design for high-throughput biomedical data. *BioMed Res. Int.* 2015;2015:958302. DOI 10.1155/2015/958302

Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*. 2022;50(W1):W345-W351. DOI 10.1093/nar/gkac247

Guizzardi G. Ontology, ontologies and the "I" of FAIR. *Data Intell*. 2020;2(1-2):181-191. DOI 10.1162/dint_a_00040

Guizzardi G., Fonseca C.M., Benevides A.B., Almeida J.P.A., Porello D., Sales T.P. Endurant Types in Ontology-Driven Conceptual Modeling: Towards OntoUML 2.0. In: Conceptual Modeling – 37th International Conference, Xi'an, China, October 22–25, 2018. Proceedings. Berlin: Springer, 2018;136-150. DOI 10.1007/978-3-030-00847-5_12

Gutierrez C., Hurtado C.A., Vaisman A. Introducing time into RDF. *IEEE Trans. Knowl. Data Eng*. 2007;19(2):207-218. DOI 10.1109/TKDE.2007.34

Hiltemann S., Rasche H., Gladman S., Hotz H.-R., Larivière D., Blankenberg D., Jagtap P.D., Wollmann T., Bretaudeau A., Goué N., Griffin T.J., Royaux C., Bras Y.L., Mehta S., Syme A., Coppens F., Droesbeke B., Soranzo N., Bacon W., Psomopoulos F., Gallardo-Alba C., Davis J., Föll M.C., Fahrner M., Doyle M.A., Serrano-Solano B., Fouilloux A.C., van Heusden P., Maier W., Clements D., Heyl F., Network G.T., Grüning B., Batut B. Galaxy Training: a powerful framework for teaching! *PLoS Comput. Biol*. 2023;19(1):e1010752. DOI 10.1371/journal.pcbi.1010752

Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N., Singhal M., Xu L., Mendes P., Kummer U. COPASI – a COmplex PAthway SImulator. *Bioinformatics*. 2006;22(24):3067-3074. DOI 10.1093/bioinformatics/btl485

Hucka M., Bergmann F.T., Chaouiya C., Dräger A., Hoops S., Keating S.M., König M., Le Novère N., Myers C.J., Olivier B.G., Sahle S., Schaff J.C., Sheriff R., Smith L.P., Waltemath D., Wilkinson D.J., Zhang F. The Systems Biology Markup Language (SBML): language specification for Level 3 Version 2 Core Release 2. *J. Integr. Bioinform*. 2019;16(2):20190021. DOI 10.1515/jib-2019-0021

Kuś M., Skowron B. (Eds.) Category Theory in Physics, Mathematics, and Philosophy, Springer Proceedings in Physics. Cham: Springer, 2019. DOI 10.1007/978-3-030-30896-4

MongoDB: The Developer Data Platform [WWW Document], n.d. MongoDB. URL https://www.mongodb.com (accessed 9.19.23)

Novère N.L., Finney A., Hucka M., Bhalla U.S., Campagne F., Collado-Vides J., Crampin E.J., Halstead M., Klipp E., Mendes P., Nielsen P., Sauro H., Shapiro B., Snoep J.L., Spence H.D., Wanner B.L. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol*. 2005;23(12):1509-1515. DOI 10.1038/nbt1156

Novère N.L., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., Wegner K., Aladjem M.I., Wimalaratne S.M., Bergman F.T., Gauges R., Ghazal P., Kawaji H., Li L., Matsuoka Y., Villéger A., Boyd S.E., Calzone L., Courtot M., Dogrusoz U., Freeman T.C., Funahashi A., Ghosh S., Jouraku A., Kim S., Kolpakov F., Luna A., Sahle S., Schmidt E., Watterson S., Wu G., Goryanin I., Kell D.B., Sander C., Sauro H., Snoep J.L., Kohn K., Kitano H. The Systems Biology Graphical Notation. *Nat. Biotechnol*. 2009;27(8):735-741. DOI 10.1038/nbt.1558

Olivier B.G., Snoep J.L. Web-based kinetic modelling using JWS Online. *Bioinformatics*. 2004;20(13):2143-2144. DOI 10.1093/bioinformatics/bth200

Petzold A., Asmi A., Vermeulen A., Pappalardo G., Bailo D., Schaap D., Glaves H.M., Bundke U., Zhao Z. ENVRI-FAIR-interoperable environmental FAIR data and services for society, innovation and research. In: 15th International Conference on eScience (eScience), San Diego, CA, USA, 2019. IEEE, 2019;277-280. DOI 10.1109/eScience.2019.00038

PostgreSQL: the world's most advanced open source database [WWW Document], n.d. URL https://www.postgresql.org/

Rad B.B., Bhatti H.J., Ahmadi M. An introduction to Docker and analysis of its performance. *Int. J. Comput. Sci. Netw. Secur.* 2017;17(3):228-235

Rocca-Serra P., Brandizi M., Maguire E., Sklyar N., Taylor C., Begley K., Field D., Harris S., Hide W., Hofmann O., Neumann S., Sterk P., Tong W., Sansone S.-A. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*. 2010;26(18):2354-2356. DOI 10.1093/bioinformatics/btq415

Roche D.G., Kruuk L.E.B., Lanfear R., Binning S.A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol*. 2015;13(11):e1002295. DOI 10.1371/journal.pbio.1002295

Schreiber F., Bader G.D., Golebiewski M., Hucka M., Kormeier B., Novère N.L., Myers C., Nickerson D., Sommer B., Waltemath D., Weise S. Specifications of standards in systems and synthetic biology. *J. Integr. Bioinform*. 2015;12(2):1-3. DOI 10.1515/jib-2015-258

Software OpenLink. Virtuoso Open-Source Edition: Building. 2022. URL https://github.com/openlink/virtuoso-opensource

Spivak D.I., Kent R.E. Ologs: a categorical framework for knowledge representation. *PLoS One*. 2012;7(1):e24274. DOI 10.1371/journal.pone.0024274

The Univalent Foundations Program. Homotopy Type Theory: Univalent Foundations of Mathematics. Princeton, NJ: Institute for Advanced Study, 2013

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.W., da Silva Santos L.B., Bourne P.E., … van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. The FAIR Guid-

ing Principles for scientific data management and stewardship. *Sci. Data*. 2016;3:160018. DOI 10.1038/sdata.2016.18

Wolstencroft K., Owen S., Krebs O., Nguyen Q., Stanford N.J., Golebiewski M., Weidemann A., Bittkowski M., An L., Shockley D., Snoep J.L., Mueller W., Goble C. SEEK: a systems biology data and model management platform. *BMC Syst. Biol*. 2015;9:33. DOI 10.1186/s12918-015-0174-y

Yan Y., Yan J. Hands-On Data Science with Anaconda: Utilize the right mix of tools to create high-performance data science applications. Packt Publishing Ltd., 2018

Zeeberg B.R., Riss J., Kane D.W., Bussey K.J., Uchio E., Linehan W.M., Barrett J.C., Weinstein J.N. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*. 2004;5:80. DOI 10.1186/1471-2105-5-80

ORCID ID

A.M. Mukhin orcid.org/0000-0002-1102-0934
F.V. Kazantsev orcid.org/0000-0002-5711-7539
S.A. Lashin orcid.org/0000-0003-3138-381X

Acknowledgements. The study is supported by the Kurchatov Genomic Centre of the Institute of Cytology and Genetics, SB RAS (No. 075-15-2019-1662).

Conflict of interest. The authors declare no conflict of interest.

Received July 13, 2023. Revised September 28, 2023. Accepted September 29, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):906-915
DOI 10.18699/VJGB-23-105

Original Russian text https://vavilovj-icg.ru/

Analysis of the transcriptional activity of model piggyBac transgenes stably integrated into different loci of the genome of CHO cells in the absence of selection pressure

L.A. Yarinich, A.A. Ogienko, A.V. Pindyurin, E.S. Omelina

Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
omelina@mcb.nsc.ru

Abstract. CHO cells are most commonly used for the synthesis of recombinant proteins in biopharmaceutical production. When stable producer cell lines are obtained, the locus of transgene integration into the genome has a great influence on the level of its expression. Therefore, the identification of genomic loci ensuring a high level of protein production is very important. Here, we used the TRIP assay to study the influence of the local chromatin environment on the activity of transgenes in CHO cells. For this purpose, reporter constructs encoding eGFP under the control of four promoters were stably integrated into the genome of CHO cells using the piggyBac transposon. Each individual transgene contained a unique tag, a DNA barcode, and the resulting polyclonal cell population was cultured for almost a month without any selection. Next, using the high-throughput sequencing, genomic localizations of barcodes, as well as their abundances in the population and transcriptional activities were identified. In total, ~640 transgenes more or less evenly distributed across all chromosomes of CHO cells were characterized. More than half of the transgenes were completely silent. The most active transgenes were identified to be inserted in gene promoters and 5′ UTRs. Transgenes carrying Chinese hamster full-length promoter of the *EF-1α* gene showed the highest activity. Transgenes with a truncated version of the same promoter and with the mouse *PGK* gene promoter were on average 10 and 19 times less active, respectively. In total, combinations of genomic loci of CHO cells and transgene promoters that together provide different levels of transcriptional activity of the model reporter construct were described.
Key words: TRIP; barcode; chromatin position effect; transgene; chromatin; transcription.

For citation: Yarinich L.A., Ogienko A.A., Pindyurin A.V., Omelina E.S. Analysis of the transcriptional activity of model piggyBac transgenes stably integrated into different loci of the genome of CHO cells in the absence of selection pressure. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):906-915. DOI 10.18699/VJGB-23-105

Анализ транскрипционной активности модельных piggyBac-трансгенов, стабильно интегрированных в разные локусы генома культивируемых клеток CHO при отсутствии селекционного давления

Л.А. Яринич, А.А. Огиенко, А.В. Пиндюрин, Е.С. Омелина

Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
omelina@mcb.nsc.ru

Аннотация. Культивируемые клетки яичника китайского хомячка (CHO) наиболее часто используются для синтеза рекомбинантных белков в биофармацевтическом производстве. При получении стабильных клеточных линий-продуцентов локус интеграции трансгена в геном оказывает большое влияние на уровень его экспрессии (явление, известное как эффект положения гена). Соответственно, поиск локусов генома, обеспечивающих высокий уровень продукции белков, является актуальной практической задачей. В данной работе мы использовали метод TRIP для исследования влияния локального окружения хроматина на активность трансгенов, встроенных в разные локусы генома культивируемых клеток CHO. С этой целью репортерные конструкции, кодирующие белок eGFP под контролем четырех разных промоторов, были стабильно встроены в геном клеток CHO при помощи транспозона piggyBac. При этом каждый отдельный трансген содержал уникальную метку – ДНК-штрихкод. Полученная трансгенная поликлональная популяция клеток была культивирована в течение месяца без какой-либо селекции. Далее при помощи присутствующих в конструкциях штрихкодов и высокопроизводительного секвенирования были определены сайты локализации трансгенов в геноме, из-

мерена их представленность в популяции, а также транскрипционная активность. Всего удалось полностью охарактеризовать около 640 трансгенов, более-менее равномерно распределенных по всем хромосомам клеток CHO. Более половины трансгенов оказались полностью молчащими. Наиболее активные трансгены выявлены в окрестностях геномных сайтов инициации транскрипции – в промоторных и 5′-некодирующих районах генов. Наибольшей активностью обладали трансгены, несущие полноразмерный промотор гена *EF-1α* китайского хомячка. Трансгены с укороченным вариантом этого же промотора, а также трансгены с промотором мышиного гена *PGK* (m*PGK*) были соответственно в среднем в 10 и 19 раз менее активны. В целом в результате данной работы выявлены сочетания локусов генома культивируемых клеток CHO и промоторных элементов, которые обеспечивают разные уровни транскрипционной активности модельной репортерной конструкции.
Ключевые слова: TRIP; штрихкод; эффект положения гена; трансген; хроматин; транскрипция.

Introduction

The TRIP (thousands of reporters integrated in parallel) assay enables large-scale studies of the influence of the chromatin position effects on the transgene activity. It is based on DNA barcodes (hereafter, barcodes) and was originally performed on the mouse embryonic stem cells using the piggyBac transposon system to deliver reporter transgenes into the genome (Akhtar et al., 2013). A barcode is a short DNA sequence (16–20 bp) that is unique to each individual transgene in the study. It should be noticed that the barcode is located within the transcribed region of the transgenes ensuring its presence not only in DNA, but also in mRNA molecules. Therefore, the barcode can be used for quantitative measurements of the level of transcriptional activity of transgenes.

The piggyBac transposon system has been previously used to effectively modify various cell lines and organisms (Wilson et al., 2007) even with large transgenic constructs (Ding et al., 2005). In addition, the piggyBac transposon is characterized by a relatively uniform distribution of insertions across chromosomes (Huang et al., 2010). In TRIP experiments, the system for transgenesis consists of two plasmid constructs: a construct for the expression of piggyBac transposase, which catalyzes the insertion of a transgene into a random genome locus, and the transgene itself – a target construct (consisting of a promoter, a reporter gene, a barcode, and a polyadenylation signal) located between the inverted terminal repeats of the piggyBac transposon (Akhtar et al., 2014; Lebedev et al., 2019). Co-transfection of cells with such plasmid constructs enables obtaining a polyclonal population of transgenic cells, in which each individual transgene insertion in the genome is marked with a unique barcode sequence. After cultivation of transfected cells, genomic DNA and total RNA are isolated from them. Using the genomic DNA sample, the genomic localization of each transgene is identified and the abundance of each barcode in the cell population is measured. Based on the total RNA sample, the abundance of each barcode in the total pool of transcripts synthesized from transgenes is measured. Finally, the ratio of the abundance of each barcode in mRNA molecules to its abundance in the cell population allows quantitative estimation of transcriptional activity of all transgenes (Akhtar et al., 2014).

In this study, we used the TRIP assay to investigate the chromatin position effects on the transcriptional activity of stably integrated transgenes in the Chinese hamster ovary cells (CHO). The CHO line is the most commonly used cell line to produce a variety of proteins (Xu et al., 2023). Despite the availability of other mammalian cells, such as baby hamster kidney cells, murine myeloma NS0 cells, human embryonic kidney cells (HEK293), human embryonic retinal PerC6 cells, more than 70 % of all recombinant therapeutic proteins are produced in CHO cells (Kim et al., 2012; Ritacco et al., 2018; Gupta et al., 2021). The popularity of CHO cells is explained by the following reasons. First, the use of CHO cells for the production of recombinant proteins is safe, since CHO cells are insensitive to infection by human viruses (Lalonde, Durocher, 2017). Second, CHO cells have the ability for efficient post-translational modification and produce recombinant proteins in human-compatible glycoforms (Stache et al., 2019). Third, CHO cells have a high growth rate and are relatively easily adapted to growth in suspension, which is a preferred characteristic for large-scale cultivation in bioreactors (Ritacco et al., 2018; Dahodwala, Lee, 2019). Currently, bioreactors with a volume of more than 10 thousand liters are used for suspension cultures of recombinant CHO cells producing therapeutic antibodies (Kim et al., 2012).

Localization in the genome has a great influence on the expression level of the recombinant gene (a phenomenon known as the chromatin position effect) (Gierman et al., 2007; Babenko et al., 2010; Ruf et al., 2011; Chen M. et al., 2013; Elgin, Reuter, 2013). Integration into inactive heterochromatin results in low or no transgene expression, whereas integration into active euchromatin often allows moderate to high transgene expression. However, simple integration into euchromatin may not be sufficient to ensure long-term expression of the recombinant gene. The phenomenon of transgene expression silencing is well known in mammalian cells, it occurs in part likely due to the influence of adjacent condensed chromatin.

Thus, integration of transgenes into transcriptionally highly active regions of the genome is a reasonable strategy to avoid position effects. This study was aimed at analyzing the transcriptional activity of the piggyBac transgenes integrated into different loci of the genome of CHO cells in the absence of selection pressure.

Materials and methods

Generation of the pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242 construct. Plasmid pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242 was made based on a previously described "universal" construct (Lebedev et al., 2019). The insertion was amplified using the primers mPGK-EcoRI-F and eGFP-XbaI-R (Table 1) and the plasmid pPTK-Gal4-mPGK-Puro-IRES-eGFP-sNRP-pA as a template (Akhtar et al., 2013). Fifty μl of the reaction mixture contained 1 ng of plasmid template, 10 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 1 μl of each 10 μM primer, 0.2 mM dNTPs, and 2.5 U of Phusion polymerase (Thermo Fisher Scientific). The PCR

Table 1. Primers used in the study

Primer name	Primer sequence (5′→3′)
mPGK-EcoRI-F	aaagaattctcgacaattctaccgggtagg
eGFP-XbaI-R	aaatctagaccctccggattacttg
hamPgk1-EcoRI-F	aaagaattcaggtccctggggattcca
hamPgk1-BglII-R	aaaagatctcggtaggatcaagaggctcag
CHEF-1-v1-EcoRI-F	aaagaattccacgttgtgcatagaaacagatgc
CHEF-1-v1-BclI-R	aaatgatcatggttttcacaacaccttaaaaaaaaagttcg
CHEF-1-v2-EcoRI-F	aaagaattcaagcttctgtggatagaaaatgattag
CHEF-1-v2-BclI-R	aaatgatcactgcgttctgacggcaaac
Plasmid-1	ccgcttaattaatccagcttttgttc
pPB-eGFP-PI-6-R	ctcgagctctcgatctctagacc
pPB-eGFP-PI-11-R	ctcactagctcgatctctagacc
pPB-eGFP-PI-16-R	ctcttgtactcgatctctagacc
pPB-eGFP-PI-28-R	ctcctcggctcgatctctagacc
PB-Barcode-PI-6-Gibson-F	gtctagagatcgagagctcgaggN_{18}gagttgtggccggcccttgtg
PB-Barcode-PI-11-Gibson-F	gtctagagatcgagctagtgaggN$_{18}$gagttgtggccggcccttgtg
PB-Barcode-PI-16-Gibson-F	gtctagagatcgagtacaagaggN$_{18}$gagttgtggccggcccttgtg
PB-Barcode-PI-28-Gibson-F	gtctagagatcgagccgaggaggN$_{18}$gagttgtggccggcccttgtg
PB-Gibson-R1	aacaaaagctggattaattaagcggccgcatacgcgtatactagattaaccc
Libr-cDNA-for	gtctcgtgggctcggagatgtgtataagagacaggtcctgctggagttcgtgac
Libr-cDNA-A16-rev	tcgtcggcagcgtcagatgtgtataagagacagcctatggtcgccagggttttcccagtcacaagg
Libr-cDNA-A23-rev	tcgtcggcagcgtcagatgtgtataagagacagtaattgcgcgccagggttttcccagtcacaagg
Libr-P5-for	aatgatacggcgaccaccgagatctacactcgtcggcagcgtc
Libr-P7-rev	caagcagaagacggcatacgagatgtctcgtgggctcgg
PB-outer-F-2	ttttacgcatgattatctttaacgtacgtc
cDNA-ampl-R	cgccagggttttcccagtcacaag
PB-cDNA-fwd-A7	tcgtcggcagcgtcagatgtgtataagagacagagcgagctgtcacaagggccggccacaa
InvPCR-F-Nextera2	gtctcgtgggctcggagatgtgtataagagacaggtacgtcacaatatgattatctttctag

Note. N_{18} – random 18-nt barcode sequence.

thermal cycle conditions were as follows: 98 °C for 30 sec, 35 cycles of 98 °C for 10 sec, 62 °C for 10 sec, 72 °C for 1 min, and a final incubation for 10 min at 72 °C.

Cloning of constructs with various Chinese hamster gene promoters. The plasmid pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242 was digested with EcoRI, BglII, and AgeI restriction enzymes. To obtain inserts, the sequences of the Chinese hamster *PGK* gene promoter and the long and short variants of the *EF-1α* gene promoter were amplified using the primers hamPgk1-EcoRI-F and hamPgk1-BglII-R, CHEF-1-v1-EcoRI-F and CHEF-1-v1-BclI-R, and CHEF-1-v2-EcoRI-F and CHEF-1-v2-BclI-R (see Table 1), respectively. Fifty μl of the reaction mixture contained 50 ng of genomic DNA template isolated from CHO cells, 10 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 1 μl of each 10 μM primer, 0.2 mM dNTPs, and 2.5 U of Phusion polymerase (Thermo Fisher Scientific). The PCR thermal cycle conditions were as follows: 98 °C for 30 sec, 35 cycles of 98 °C for 10 sec, 62 °C for 10 sec, 72 °C for 1 min, and a final incubation for 10 min at 72 °C.

Generation of barcoded plasmid libraries. Barcoded plasmid libraries were made according to a previously described protocol (Lebedev et al., 2019) using the Gibson cloning method. For this purpose, vectors and inserts containing an

18-nt DNA barcode and a promoter index were prepared using PCR. For vector amplification, the primers Plasmid-1 and pPB-eGFP-PI-6-R / pPB-eGFP-PI-11-R / pPB-eGFP-PI-16-R / pPB-eGFP-PI-28-R were used (see Table 1) for the constructs with the Chinese hamster *PGK* gene promoter / m*PGK* gene promoter / short variant of the *EF-1α* gene promoter / long variant of the *EF-1α* gene promoter, respectively. The primers PB-Barcode-PI-6-Gibson-F / PB-Barcode-PI-11-Gibson-F / PB-Barcode-PI-16-Gibson-F / PB-Barcode-PI-28-Gibson-F and PB-Gibson-R1 (see Table 1) were used to amplify the barcoded inserts for constructs with the Chinese hamster *PGK* gene promoter / the *mPGK* gene promoter / short variant of the *EF-1α* gene promoter / long variant of the *EF-1α* gene promoter, respectively. Fifty μl of the reaction mixture contained 1 ng of template, 10 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 1 μl of each 10 μM primer, 0.2 mM dNTPs, and 2.5 U of Phusion polymerase (Thermo Fisher Scientific). The PCR thermal cycle conditions were as follows: 98 °C for 30 sec, 35 cycles of 98 °C for 10 sec, 62 °C for 10 sec, 72 °C for 1 min, and a final incubation for 10 min at 72 °C. After purification, 200 ng of "vector" and 135 ng of "inserts" were mixed with 10 μl of 2× NEBuilder HiFi DNA Assembly Master in a total volume of 20 μl. DNA ligation and bacterial transformation were performed as described previously (Lebedev et al., 2019). Barcoded plasmid libraries were isolated using the Mega Plasmid Kit (Qiagen).

Generation of polyclonal transgenic population of CHO cells. Twenty-four h before transfection, CHO-S cells (hereafter CHO cells; kindly provided by Dr. A.V. Taranin, Institute of Molecular and Cellular Biology, Novosibirsk, Russia) were seeded into a 12-well culture plate at a concentration of 1.5×10^5 cells per ml in IMDM medium supplemented with 10 % bovine serum. Cells were co-transfected with a mixture of barcoded plasmid libraries (3 μg) and the pRP[Exp]-mCherry-CAG>hyPBase plasmid (VectorBuilder #VB160216-10057; kindly provided by Prof. V.V. Verkhusha, Albert Einstein College of Medicine, Bronx, NY, USA) (0.3 μg) using the X-tremeGENE HP DNA transfection reagent (Roche). The transfected cells were cultured for a month in the absence of selection pressure.

Isolation of genomic DNA. Genomic DNA was isolated from $5 \cdot 10^7$ cells of the resulting polyclonal transgenic population using the PureLink® Genomic DNA Kit (Invitrogen) according to the manufacturer's recommendations.

Isolation of total RNA, reverse transcription. Total RNA was isolated from $5 \cdot 10^7$ cells of the resulting polyclonal transgenic population using RNAzol RT (Molecular Research Center) according to the manufacturer's recommendations. The isolated RNA was incubated with 20 U of DpnI restriction endonuclease (New England Biolabs) and 3 U of DNase I (Thermo Fisher Scientific) for 30 min at 37 °C. The Clean RNA Standard kit (Evrogen) was used to purify RNA. Two μg of purified total RNA was mixed with 1 μl of the 50 mM oligo(dT) primer in a total volume of 13.5 μl, and the mixture was incubated for 5 min at 65 °C. The subsequent reverse transcription reaction was carried out in a volume of 20 μl with the following components: 13.5 μl of RNA with annealed primer, 4 μl of 5× RT buffer (Thermo Fisher Scientific), 1 μl of 10 mM dNTPs, 1 μl of RNaseOUT (Thermo Fisher Scientific), 100 U of RevertAid reverse transcriptase (Thermo Fisher Scientific). The mixture was incubated for 60 min at 42 °C, and the enzyme was inactivated for 10 min at 70 °C.

Preparation of the normalization and expression samples. To prepare each sample, two rounds of PCR were performed. For the first round of amplification, we used 600 ng of genomic DNA template (for the normalization sample) or 3 μl of cDNA (for the expression sample), 0.5 μl of 10 μM primers Libr-cDNA-for and Libr-cDNA-A16-rev / Libr-cDNA-A23-rev (see Table 1) for the normalization / expression sample, respectively, 5 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 2 μl of 2.5 mM dNTPs, and 1.25 U of Phusion HS II DNA polymerase (Thermo Fisher Scientific) in a total volume of 25 μl. The thermal cycle conditions of the first round of PCR were as follows: 98 °C for 1 min, 15 cycles of 98 °C for 30 sec, 70 °C for 30 sec, 72 °C for 30 sec, and a final incubation for 5 min at 72 °C. The second round of amplification was carried out in a volume of 25 μl with the following components: 0.5 μl of the PCR products of the first round, 0.25 μl of 10 μM primers Libr-P5-for and Libr-P7-rev (see Table 1), 5 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 2 μl of 2.5 mM dNTPs, and 1.25 U of Phusion HotStart II DNA polymerase (Thermo Fisher Scientific). The thermal cycle conditions of the second round of PCR were as follows: 98 °C for 1 min, 23 cycles of 98 °C for 30 sec, 61 °C for 30 sec, 72 °C for 30 sec, and a final incubation for 5 min at 72 °C.

Preparation of the mapping sample. Two μg of genomic DNA was incubated with 10 U of DpnII restriction endonuclease (New England Biolabs) at 37 °C for 16 h, then purified using the GeneJET PCR Purification Kit (Thermo Fisher Scientific). 600 ng of the digested genomic DNA was mixed with 4 μl of 100 mM ATP, 2.5 U of T4 DNA ligase (Evrogen) in a total volume of 400 μl. The ligase mixture was incubated for 2 h at room temperature and then for 16 h at 4 °C, followed by enzyme inactivation at 65 °C for 10 min. 100 μl of double-distilled water and 500 μl of a phenol:chloroform solution (1:1 ratio) were added to the ligation reaction, mixed, centrifuged at room temperature for 5 min at 10,000×*g*, and the upper phase was transferred to a new tube. One-tenth volume of 3M NaOAc (pH 5.5) and 2.5 volumes of 96 % ethyl alcohol were added to the resulting solution, the mixture was incubated for 2 h at –70 °C, then centrifuged for 30 min at 4 °C, 14,000 rpm. The supernatant was removed, the pellet was washed with 750 μl of chilled 70 % ethyl alcohol, and centrifuged for 10 min at 4 °C, 14,000 rpm. The supernatant was removed, the pellet was dried for 15 min at 37 °C and then dissolved in 30 μl of double-distilled water.

Three rounds of PCR were used to prepare the mapping sample. For the first round of amplification, we used 5 μl of purified ligase mixture, 0.5 μl of 10 μM primers PB-outer-F-2 and cDNA-ampl-R (see Table 1), 5 μl of 5× Phusion HF buffer (Thermo Fisher Scientific), 2 μl of 2.5 mM dNTPs, and 1.25 U of Phusion HS II DNA polymerase (Thermo Fisher Scientific) in a total volume of 25 μl. The subsequent rounds of amplification were carried out similarly using (i) primers PB-cDNA-fwd-A7 and InvPCR-F-Nextera2 (see Table 1) and 1 μl of the PCR products of the first round for the second round

of amplification and (ii) primers Libr-P5-for and Libr-P7-rev (see Table 1) and 1 μl of the PCR products of the second round for the third round of amplification. The thermal cycle conditions for the first round of PCR were as follows: 98 °C for 1 min, 12 cycles of 98 °C for 30 sec, 65 °C for 30 sec, 72 °C for 2 min, and a final incubation for 5 min at 72 °C. The thermal cycle conditions for the second round of PCR were as follows: 98 °C for 1 min, 12 cycles of 98 °C for 30 sec, 62 °C for 30 sec, 72 °C for 2 min, and a final incubation for 5 min at 72 °C. The thermal cycle conditions for the third round of PCR were as follows: 98 °C for 1 min, 16 cycles of 98 °C for 30 sec, 61 °C for 30 sec, 72 °C for 2 min, and a final incubation for 5 min at 72 °C. Finally, 5 μg of the mapping sample was treated with 10 U of NotI restriction endonuclease (New England Biolabs) in a total volume of 100 μl at 37 °C for 2 h to remove byproducts.

High-throughput sequencing and data analysis. Sequencing of the samples was performed on the Genolab 2 × 75 bp platform (https://genomed.ru/). Demultiplexing of the obtained fastq files using the sabre tool (https://github.com/najoshi/sabre) resulted in ~4.5, ~1.6 and ~1 million reads for the mapping, normalization, and expression samples, respectively. The quality analysis of the raw data was carried out using the FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Next, using the TASK tool (The TRIP Analysis Software Kit, https://trip.nki.nl/), the sequences of reliably identified barcodes, as well as their normalized expression levels and genomic locations in the Chinese hamster genome assemblies CriGri-PICRH-1.0 (GCA_003668045.2) and Cgr1.0 (GCA_000448345.1) were identified. The genome assembly CriGri-PICRH-1.0 is characterized by the presence of extremely long sequences corresponding to all expected chromosomes, and therefore it was used as the "default" genome of CHO cells, whereas the genome assembly Cgr1.0 was previously used to map chromatin types in CHO-K1 cells (Feichtinger et al., 2016). Then, to determine the most reliable (hereafter referred to as filtered) transgenes, the following additional parameters were applied to the TASK output: norm >= 5, reads_r >= 10, freq1_r > 0.60. Data on chromatin types were taken for the Tp0 time point corresponding to the culturing of CHO-K1 cells for 4 h (https://cho-epigenome.boku.ac.at/JB/). A positional weight matrix for genomic sequences overlapping transgene insertion sites was generated using the pLogo application (https://plogo.uconn.edu/) (O'Shea et al., 2013).

Results and discussion

To study the activity of several promoters in different local chromatin environments in CHO cells in parallel, barcoded model transgenes were constructed based on the piggyBac transposon. The transgenes contained the puromycin-*N*-acetyltransferase (*pac*) resistance gene (hereafter $Puro^R$) and the improved green fluorescent protein (e*GFP*) gene (separated by an IRES element) under the control of the following four promoters: (1) the promoter of the mouse *PGK* (m*PGK*) gene, previously used for a similar study on mouse embryonic stem cells (Akhtar et al., 2013), (2) the promoter of the Chinese hamster *PGK* gene, homologous to the m*PGK* promoter, (3) full-length ("long") and (4) truncated ("short") variants of the Chinese hamster *EF-1α* gene promoter (Running Deer, Allison, 2004; Orlova et al., 2014; Wang et al., 2017) (Fig. 1, *A*). Constructs with each individual promoter also contained a specific 5-bp motif (promoter index) immediately before the 18-bp barcode located in the 3′UTR region. The presence of promoter indexes allowed the simultaneous use of all 4 barcoded model transgenes in a single experiment (Gisler et al., 2019) (see Fig. 1, *A*). Thus, the resulting barcoded plasmid libraries of the constructs with the long *EF-1α,* short *EF-1α*, m*PGK* and *PGK* promoters were mixed in a molar ratio of 7:7:7:1. The smaller proportion of the latter construct was due to its use in this experiment as a control; we also used the *PGK* promoter-containing construct to obtain stable transgenic populations of CHO cells upon puromycin selection (the results of that study will be reported elsewhere) and it seemed useful to us to have the technical ability to correctly compare data for such different transgenic populations in the future.

CHO cells (CHO-S subline) were co-transfected with the above-described mixture of model transgenes, as well as a plasmid encoding the piggyBac transposase. Seventy-two h after transfection, the eGFP protein expression was observed in approximately 40 % of the cells (see Fig. 1, *B*). After that, the cells were cultivated in the absence of any selection for additional 25 days in order to multiply the transgenic cells and to get rid of plasmid DNA molecules that could ultimately contaminate the data. As a result, multiple clones of transgenic cells were observed in the population (see Fig. 1, *B*).

From the resulting polyclonal population of cells, genomic DNA and total RNA were isolated, on the basis of which the genomic localizations and normalized expression levels of barcoded transgenes were determined. A total of 641 uniquely barcoded and genome-mapped transgenes were identified in the transgenic population. These transgenes were present in more or less expected numbers on all chromosomes of CHO cells (see Fig. 1, *C*). Analysis of genomic sequences overlapping transgene insertion sites revealed their at-richness, as well as the presence of a central ttaa motif (see Fig. 1, *D*) specific for the piggyBac transposon (Frase et al., 1996; Li et al., 2013; Chen Q. et al., 2020).

Among the identified transgenes, 38.8 % were with the tacaa promoter index (corresponding to the short *EF-1α* promoter), 24.3 %, with the ccgag promoter index (corresponding to the long *EF-1α* promoter), 32.2 %, with the ctagt promoter index (corresponding to the m*PGK* promoter), and 4.7 %, with the promoter index agctc (corresponding to the Chinese hamster *PGK* promoter) (Fig. 2, *A*).

Analysis of the activity of reporter constructs under the control of four different promoters revealed the presence of a large number of silent (i. e., transcriptionally inactive) transgenes with each promoter (Table 2), which is most likely due to the lack of antibiotic selection during generation of the population of transgenic cells.

Comparison of promoter activities among the filtered expressed transgenes (144 cases) showed that the main part of highly active reporter constructs is under the control of the full-length variant of the *EF-1α* gene promoter (see Fig. 2, *B,* Table 2). Particularly, among 10 % of the most active filtered transgenes, the promoters were distributed as follows: long *EF-1α* – 70 %, m*PGK* – 20 %, short *EF-1α* – 10 %, *PGK* – 0 %.

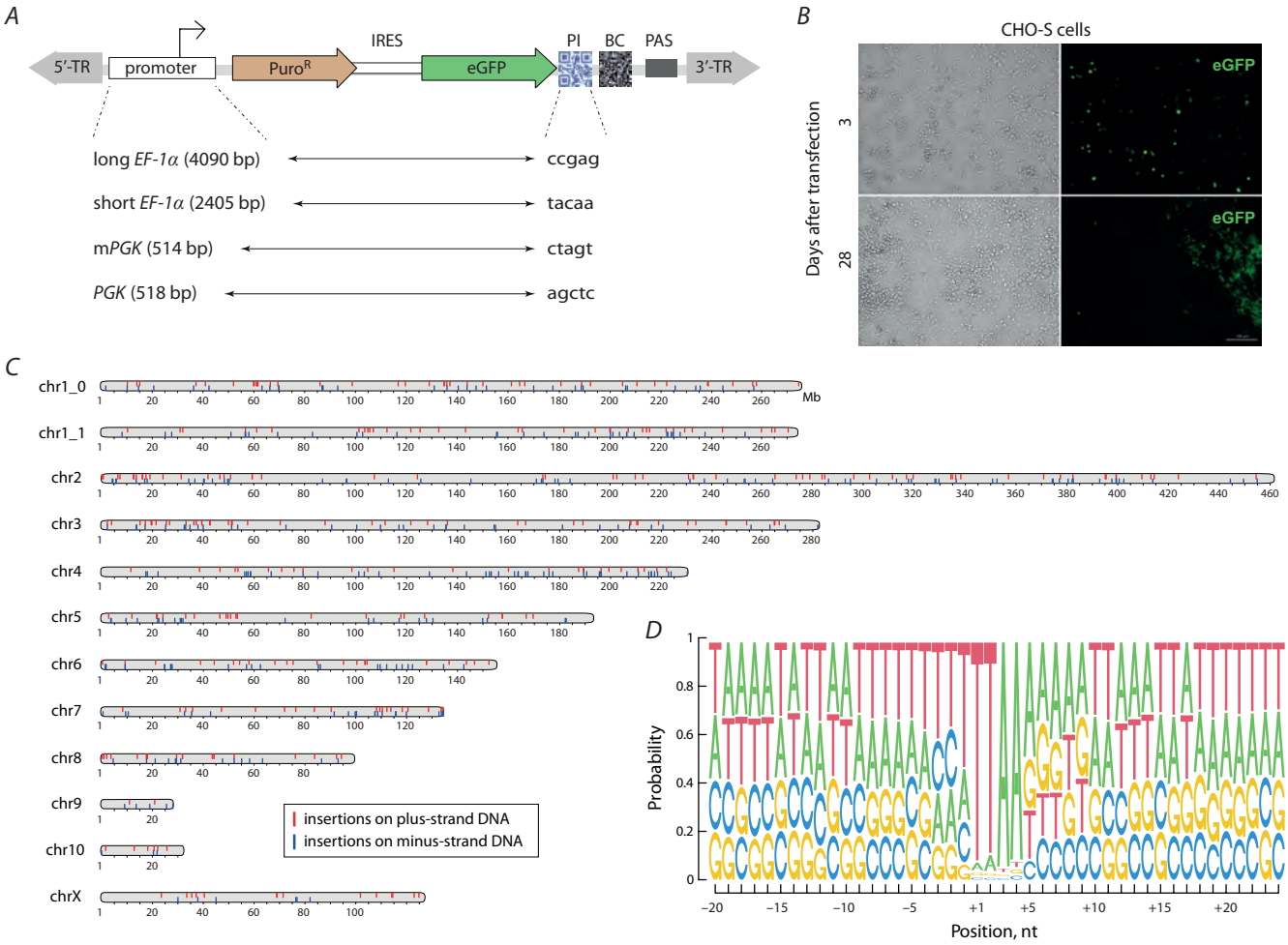


Fig. 1. Stable integration of model piggyBac transgenes into the genome of cultured CHO cells.

A, Schematic of the barcoded reporter constructs with four different promoters used in the study. 5′-TR and 3′-TR are inverse terminal repeats of the piggyBac transposon; IRES – internal ribosome entry site, PI – promoter index, BC – barcode, PAS – polyadenylation signal. *B*, CHO cells 3 and 28 days after transfection. *C*, Distribution of all uniquely mapped transgenes across Chinese hamster chromosomes. Red and blue vertical lines indicate transgene insertions on the plus and minus strand of the genome, respectively. *D*, Analysis of the genomic motifs, at which integration of all uniquely mapped transgenes occurred. Positions +1…+4 correspond to the sequence that is duplicated upon the piggyBac transposon insertion and flanks the integrated transgene.

It is worth noting that due to the small number of transgenes with the *PGK* promoter, the results on its activity are rather preliminary.

Two thirds of all transgenes were inserted into the genome of CHO cells within genes (which were defined as –1000 bp from the distal transcription initiation site to the transcription termination site), predominantly in introns (42.3 %), promoters (8.5 %) and 5′UTRs (9.4 %) (see Fig. 2, *C*). It should be noted that promoters were defined as –1000…+100 bp from the transcription start site. Similar patterns of transgene integration based on the piggyBac transposon have been described previously for cell cultures of other species (Ding et al., 2005; Wilson et al., 2007; Galvan et al., 2009; Li et al., 2013). Analysis of 10 % of the most active filtered reporter constructs (21 cases) revealed a ~1.6-fold, ~1.4-fold and ~1.1-fold increase in the proportion of transgenes in 5′UTRs, promoters and introns, respectively (see Fig. 2, *D*). In addition, it is interesting to note that transgenes were more often located closer to the gene starts than to the gene ends. The median distances from the transgene localization position in the genome to the nearest transcription initiation and termination sites were 11.4 and 20.2 kb, respectively, for the complete set of reporter genes (641 cases). At the same time, for 10 % of the most active filtered transgenes (21 cases), these values were equal to 6.6 and 17.8 kb, respectively.

To study the influence of the local chromatin environment on the activity of reporter genes, previously published data on the distribution of 11 chromatin types in the genome of CHO-K1 cells were used (Feichtinger et al., 2016). Because these data were originally associated with a version of the Chinese hamster genome that is different from that used for the analysis described above (see Materials and methods for details), it was possible to extract chromatin types overlapping positions of transgenes in the genome only for 595 out of 641 transgenes. Among these 595 transgenes, only 39.5 % were localized in inactive chromatin types “Quiescent/low”, “Repressed heterochromatin (H3K9me3)” and “Polycomb repressed regions (H3K27me3)”, which together cover more

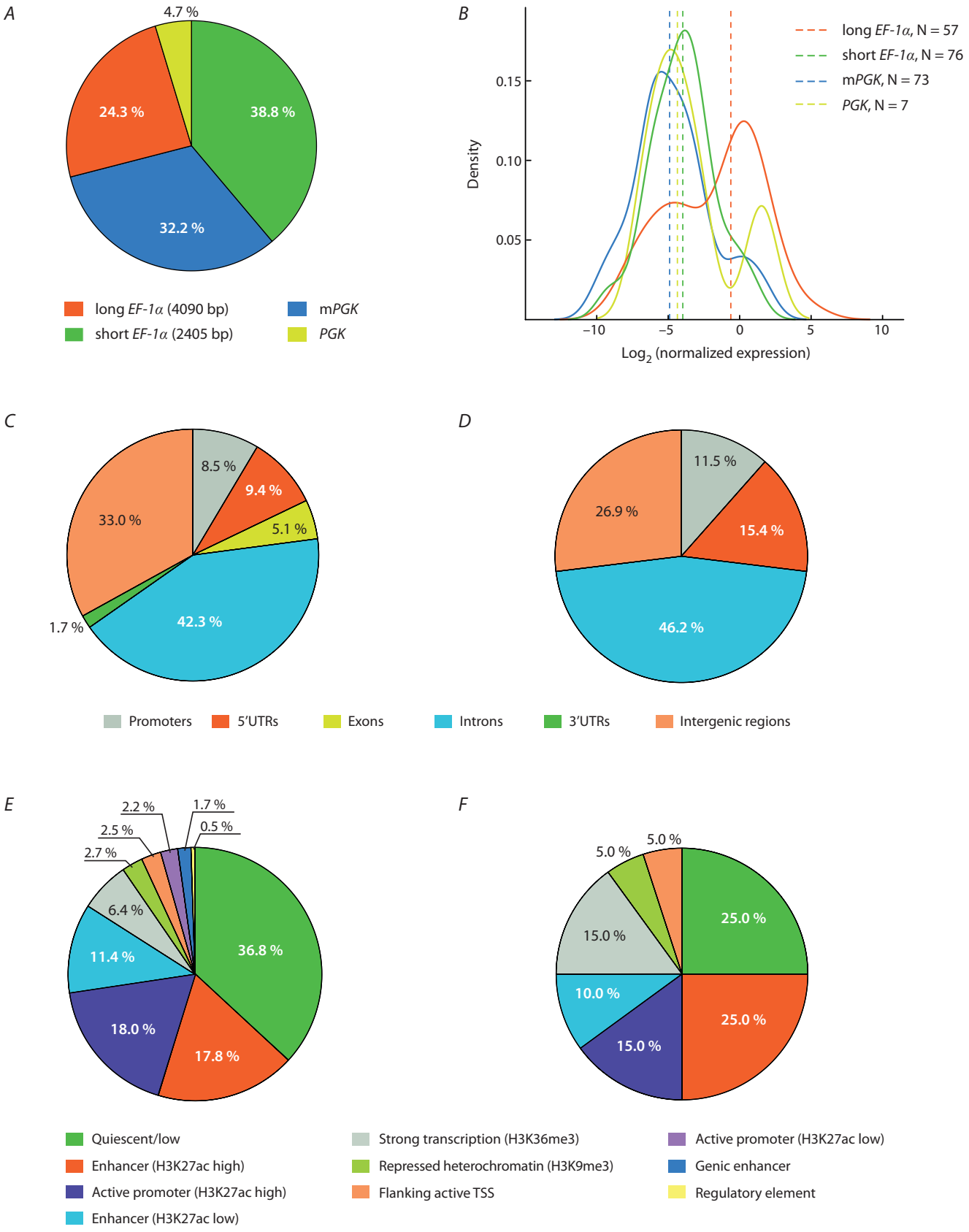


Fig. 2. Features of studied transgenes.

A, Distribution of all identified transgenes across the studied promoters. *B*, Comparison of promoter activities for the filtered 144 expressed transgenes (see Materials and methods). Dashed vertical lines indicate median normalized expression values. *C*, *D*, Distribution of all identified transgenes (*C*) and 10 % of the most active filtered transgenes (*D*) in gene elements (promoters, 5'UTRs, exons, introns, 3'UTRs) as well as intergenic regions. *E*, *F*, Distribution of all transgenes (*E*) and 10 % of the most active filtered transgenes (*F*) in chromatin types of CHO-K1 cells described previously (Feichtinger et al., 2016).

Table 2. Comparative activity of the studied promoters

Promoter	Transgene number	Proportion of silent transgenes, %	Proportion of active transgenes, %	Median promoter activity value, a. u.*
long *EF-1α*	156	42.31	57.69	18.93
short *EF-1α*	249	60.24	39.76	1.84
m*PGK*	206	53.88	46.12	1
PGK	30	66.67	33.33	1.42
Total	641	54.13	45.87	

* a. u. – arbitrary units.

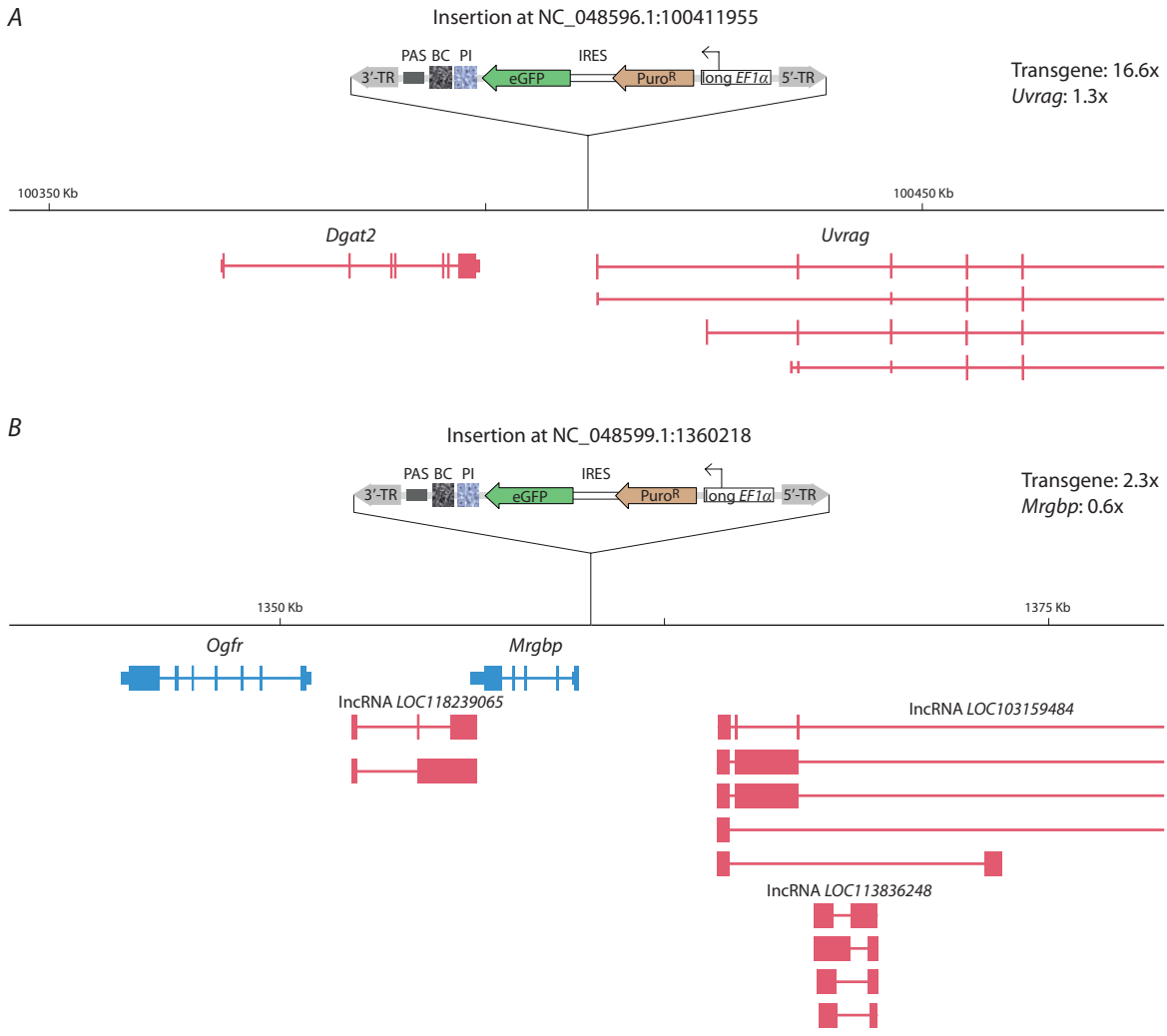


Fig. 3. Examples of genomic localization of transgenes with high (*A*) and medium (*B*) transcriptional activity.

The transgenes are not shown in scale. The activities of the transgenes and the genes nearest to them are indicated relative to the average expression levels of all reporter genes (641 cases) and all endogenous genes, respectively.

than 88 % of the genome of Chinese hamster cells (Feichtinger et al., 2016). The remaining 60.5 % of transgenes were localized in various active chromatin types (see Fig. 2, *E*).

The most active transgenes were more often localized in regions of the genome associated with the active chromatin types “Enhancer (H3K27ac high)”, “Strong transcription (H3K36me3)” and “Flanking active TSS” as well as with the inactive chromatin type “Repressed heterochromatin (H3K9me3)” (see Fig. 2, *F*). The latter rather unexpected observation may be due to the fact that the chromatin types were determined for a different subline of CHO cells.

Since, as noted above, two thirds of all transgenes were localized within genes (see Fig. 2, *C*), it is worth noting that the insertion of a transgene even into an important gene most

likely has only a minor effect on cell viability. This is supported by the following two considerations. First, not every insertion of a foreign sequence within a gene significantly disrupts its function. Second, typically, there is another native copy(ies) of the gene in the genome of cultured cells. Together, this ensures the successful survival of transgenic cells in a polyclonal population among non-modified (wild-type) cells. The chances of damaging both alleles of a gene in the experimental setup used are negligible: to achieve that, two transgenes must be integrated into both alleles of the same gene in the same cell. Accordingly, the genomic positions of active transgenes identified in this study may qualify for consideration as promising sites for targeted integration of biotechnological transgenes, even if they are located inside active genes (Fig. 3).

Conclusion

In a polyclonal population of transgenic CHO cells cultured in the absence of selection pressure, more than half of the model reporter constructs stably integrated into the genome were transcriptionally inactive. Compared to the complete set of transgenes, the most active transgenes were about 1.6 and 1.4 times more frequently localized in promoters and 5′UTRs of genes, respectively. Also, compared to the complete set of transgenes, the most active transgenes were about 2.3 and 1.4 times more frequently localized in the transcriptionally active chromatin types “Strong transcription (H3K36me3)” and “Enhancer (H3K27ac high)”, respectively. Transgenes containing the full-length promoter of the Chinese hamster *EF-1α* gene were on average the most active ones. At the same time, the median activity of the short variant of the *EF-1α* gene promoter was about 10 times lower than the median activity of the full-length promoter of this gene. This can be explained by the presence of important binding sites for transcription factors in the full-length version of the *EF-1α* gene promoter. Genomic sites of the most active insertions of model transgenes may be of interest for further experiments as promising positions for targeted integration of biotechnological constructs.

References

Akhtar W., de Jong J., Pindyurin A.V., Pagie L., Meuleman W., de Ridder J., Berns A., Wessels L.F.A., van Lohuizen M., van Steensel B. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*. 2013;154(4):914-927. DOI 10.1016/j.cell.2013.07.018

Akhtar W., Pindyurin A.V., de Jong J., Pagie L., ten Hoeve J., Berns A., Wessels L.F.A., van Steensel B., van Lohuizen M. Using TRIP for genome-wide position effect analysis in cultured cells. *Nat. Protoc*. 2014;9(6):1255-1281. DOI 10.1038/nprot.2014.072

Babenko V.N., Makunin I.V., Brusentsova I.V., Belyaeva E.S., Maksimov D.A., Belyakin S.N., Maroy P., Vasil′eva L.A., Zhimulev I.F. Paucity and preferential suppression of transgenes in late replication domains of the *D. melanogaster* genome. *BMC Genomics*. 2010;11:318. DOI 10.1186/1471-2164-11-318

Chen M., Licon K., Otsuka R., Pillus L., Ideker T. Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep.* 2013;3(1):128-137. DOI 10.1016/j.celrep.2012.12.003

Chen Q., Luo W., Veach R.A., Hickman A.B., Wilson M.H., Dyda F. Structural basis of seamless excision and specific targeting by *piggyBac* transposase. *Nat. Commun*. 2020;11(1):3446. DOI 10.1038/s41467-020-17128-1

Dahodwala H., Lee K.H. The fickle CHO: a review of the causes, implications, and potential alleviation of the CHO cell line instability problem. *Curr. Opin. Biotechnol*. 2019;60:128-137. DOI 10.1016/j.copbio.2019.01.011

Ding S., Wu X., Li G., Han M., Zhuang Y., Xu T. Efficient transposition of the *piggyBac* (*PB*) transposon in mammalian cells and mice. *Cell*. 2005;122(3):473-483. DOI 10.1016/j.cell.2005.07.013

Elgin S.C.R., Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb. Perspect. Biol*. 2013;5(8):a017780. DOI 10.1101/cshperspect.a017780

Feichtinger J., Hernández I., Fischer C., Hanscho M., Auer N., Hackl M., Jadhav V., Baumann M., Krempl P.M., Schmidl C., Farlik M., Schuster M., Merkel A., Sommer A., Heath S., Rico D., Bock C., Thallinger G.G., Borth N. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. *Biotechnol. Bioeng*. 2016;113(10):2241-2253. DOI 10.1002/bit.25990

Fraser M.J., Ciszczon T., Elick T., Bauser C. Precise excision of TTAA-specific lepidopteran transposons *piggyBac* (IFP2) and *tagalong* (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol. Biol*. 1996;5(2):141-151. DOI 10.1111/j.1365-2583.1996.tb00048.x

Galvan D.L., Nakazawa Y., Kaja A., Kettlun C., Cooper L.J.N., Rooney C.M., Wilson M.H. Genome-wide mapping of *PiggyBac* transposon integrations in primary human T cells. *J. Immunother*. 2009;32(8):837-844. DOI 10.1097/CJI.0b013e3181b2914c

Gierman H.J., Indemans M.H.G., Koster J., Goetze S., Seppen J., Geerts D., van Driel R., Versteeg R. Domain-wide regulation of gene expression in the human genome. *Genome Res*. 2007;17(9):1286-1295. DOI 10.1101/gr.6276007

Gisler S., Gonçalves J.P., Akhtar W., de Jong J., Pindyurin A.V., Wessels L.F.A., van Lohuizen M. Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nat. Commun*. 2019;10(1):1598. DOI 10.1038/s41467-019-09551-w

Gupta K., Modi D., Jain R., Dandekar P. A stable CHO K1 cell line for producing recombinant monoclonal antibody against TNF-α. *Mol. Biotechnol*. 2021;63(9):828-839. DOI 10.1007/s12033-021-00329-4

Huang X., Guo H., Tammana S., Jung Y.-C., Mellgren E., Bassi P., Cao Q., Tu Z.J., Kim Y.C., Ekker S.C., Wu X., Wang S.M., Zhou X. Gene transfer efficiency and genome-wide integration profiling of *Sleeping Beauty*, *Tol2*, and *piggyBac* transposons in human primary T cells. *Mol. Ther*. 2010;18(10):1803-1813. DOI 10.1038/mt.2010.141

Kim J.Y., Kim Y.-G., Lee G.M. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Appl. Microbiol. Biotechnol*. 2012;93(3):917-930. DOI 10.1007/s00253-011-3758-5

Lalonde M.-E., Durocher Y. Therapeutic glycoprotein production in mammalian cells. *J. Biotechnol*. 2017;251:128-140. DOI 10.1016/j.jbiotec.2017.04.028

Lebedev M.O., Yarinich L.A., Ivankin A.V., Pindyurin A.V. Generation of barcoded plasmid libraries for massively parallel analysis of chromatin position effects. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2019;23(2):203-211. DOI 10.18699/VJ19.483

Li M.A., Pettitt S.J., Eckert S., Ning Z., Rice S., Cadiñanos J., Yusa K., Conte N., Bradley A. The *piggyBac* transposon displays local and distant reintegration preferences and can cause mutations at non-canonical integration sites. *Mol. Cell. Biol.* 2013;33(7):1317-1330. DOI 10.1128/MCB.00670-12

Orlova N.A., Kovnir S.V., Hodak J.A., Vorobiev I.I., Gabibov A.G., Skryabin K.G. Improved elongation factor-1 alpha-based vectors for stable high-level expression of heterologous proteins in Chinese hamster ovary cells. *BMC Biotechnol.* 2014;14:56. DOI 10.1186/1472-6750-14-56

O'Shea J.P., Chou M.F., Quader S.A., Ryan J.K., Church G.M., Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods*. 2013;10(12):1211-1212. DOI 10.1038/nmeth.2646

Ritacco F.V., Wu Y., Khetan A. Cell culture media for recombinant protein expression in Chinese hamster ovary (CHO) cells: history, key components, and optimization strategies. *Biotechnol. Prog*. 2018;34(6):1407-1426. DOI 10.1002/btpr.2706

Ruf S., Symmons O., Uslu V.V., Dolle D., Hot C., Ettwiller L., Spitz F. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.* 2011;43(4):379-386. DOI 10.1038/ng.790

Running Deer J., Allison D.S. High-level expression of proteins in mammalian cells using transcription regulatory sequences from the Chinese hamster EF-1α gene. *Biotechnol. Prog*. 2004;20(3):880-889. DOI 10.1021/bp034383r

Stach C.S., McCann M.G., O'Brien C.M., Le T.S., Somia N., Chen X., Lee K., Fu H.Y., Daoutidis P., Zhao L., Hu W.S., Smanski M. Model-driven engineering of N-linked glycosylation in Chinese hamster ovary cells. *ACS Synth. Biol*. 2019;8(11):2524-2535. DOI 10.1021/acssynbio.9b00215

Wang X., Xu Z., Tian Z., Zhang X., Xu D., Li Q., Zhang J., Wang T. The EF-1α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells. *J. Cell. Mol. Med*. 2017;21(11):3044-3054. DOI 10.1111/jcmm.13216

Wilson M.H., Coates C.J., George A.L., Jr. *PiggyBac* transposon-mediated gene transfer in human cells. *Mol. Ther*. 2007;15(1):139-145. DOI 10.1038/sj.mt.6300028

Xu W.-J., Lin Y., Mi C.-L., Pang J.-Y., Wang T.-Y. Progress in fed-batch culture for recombinant protein production in CHO cells. *Appl. Microbiol. Biotechnol*. 2023;107(4):1063-1075. DOI 10.1007/s00253-022-12342-x

ORCID ID

L.A. Yarinich orcid.org/0000-0003-0469-0371
A.A. Ogienko orcid.org/0000-0002-0896-1899
A.V. Pindyurin orcid.org/0000-0001-6959-0641
E.S. Omelina orcid.org/0000-0002-2189-5101

Acknowledgements. The work was carried out with financial support from the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2021-1086, contract RF----193021X0015, 15.IP.21.0015).
We are grateful to A.V. Taranin (Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia), V.V. Verkhusha (Albert Einstein College of Medicine, Bronx, NY, USA) and V.S. Fishman (Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia) for providing CHO-S cells, the pRP[Exp]-mCherry-CAG>hyPBase plasmid and assistance in extracting data on the chromatin types of CHO cells, respectively.

Conflict of interest. The authors declare no conflict of interest.

Received September 13, 2023. Revised September 21, 2023. Accepted September 27, 2023.

Прием статей через электронную редакцию на сайте http://vavilov.elpub.ru/index.php/jour
Предварительно нужно зарегистрироваться как автору, затем в правом верхнем углу страницы выбрать «Отправить рукопись». После завершения загрузки материалов обязательно выбрать опцию «Отправить письмо», в этом случае редакция автоматически будет уведомлена о получении новой рукописи.

«Вавиловский журнал генетики и селекции»/"Vavilov Journal of Genetics and Breeding"
до 2011 г. выходил под названием «Информационный вестник ВОГиС»/
"The Herald of Vavilov Society for Geneticists and Breeding Scientists".

Сетевое издание «Вавиловский журнал генетики и селекции» – реестровая запись СМИ
Эл № ФС77-85772, зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 14 августа 2023 г.

Издание включено ВАК Минобрнауки России в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, Russian Science Citation Index на платформе Web of Science, Российский индекс научного цитирования, ВИНИТИ, Web of Science CC, Scopus, PubMed Central, DOAJ, ROAD, Ulrich's Periodicals Directory, Google Scholar.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте ИЦиГ СО РАН, https://vavilovj-icg.ru/
и платформе Научной электронной библиотеки, elibrary.ru/title_about.asp?id=32440

англоязычная версия – на сайте vavilov.elpub.ru/index.php/jour
и платформе PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/journals/3805/

✉ email: vavilov_journal@bionet.nsc.ru

Издатель: Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук»,
проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Адрес редакции: проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Секретарь по организационным вопросам С.В. Зубова. Тел.: (383)3634977.
Издание подготовлено информационно-издательским отделом ИЦиГ СО РАН. Тел.: (383)3634963*5218.
Начальник отдела: Т.Ф. Чалкова. Редакторы: В.Д. Ахметова, И.Ю. Ануфриева. Дизайн: А.В. Харкевич.
Компьютерная графика и верстка: Т.Б. Коняхина, О.Н. Савватеева.
Фотография на обложке О.В. Андреенкова.