

Научный рецензируемый журнал

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.

Периодичность 8 выпусков в год

DOI 10.18699/VJGB-22-86

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

Главный редактор

А.В. Кочетов – академик РАН, д-р биол. наук (Россия)

Заместители главного редактора

Н.А. Колчанов – академик РАН, д-р биол. наук, профессор (Россия)

И.Н. Леонова – д-р биол. наук (Россия)

Н.Б. Рубцов – д-р биол. наук, профессор (Россия)

В.К. Шумный – академик РАН, д-р биол. наук, профессор (Россия)

Ответственный секретарь

Г.В. Орлова – канд. биол. наук (Россия)

Редакционная коллегия

Е.Е. Андронов – канд. биол. наук (Россия)

Ю.С. Аульченко – д-р биол. наук (Россия)

О.С. Афанасенко – академик РАН, д-р биол. наук (Россия)

Д.А. Афонников – канд. биол. наук, доцент (Россия)

Л.И. Афтанас – академик РАН, д-р мед. наук (Россия)

Л.А. Беспалова – академик РАН, д-р с.-х. наук (Россия)

А. Бёрнер – д-р наук (Германия)

Н.П. Бондарь – канд. биол. наук (Россия)

С.А. Боринская – д-р биол. наук (Россия)

П.М. Бородин – д-р биол. наук, проф. (Россия)

А.В. Васильев – чл.-кор. РАН, д-р биол. наук (Россия)

М.И. Воевода – академик РАН, д-р мед. наук (Россия)

Т.А. Гавриленко – д-р биол. наук (Россия)

И. Гроссе – д-р наук, проф. (Германия)

Н.Е. Грунтенко – д-р биол. наук (Россия)

С.А. Демаков – д-р биол. наук (Россия)

И.К. Захаров – д-р биол. наук, проф. (Россия)

И.А. Захаров-Гезехус – чл.-кор. РАН, д-р биол. наук (Россия)

С.Г. Инге-Вечтомов – академик РАН, д-р биол. наук (Россия)

А.В. Кильчевский – чл.-кор. НАНБ, д-р биол. наук (Беларусь)

С.В. Костров – чл.-кор. РАН, д-р хим. наук (Россия)

А.М. Кудрявцев – чл.-кор. РАН, д-р биол. наук (Россия)

И.Н. Лаврик – д-р биол. наук (Германия)

Д.М. Ларкин – канд. биол. наук (Великобритания)

Ж. Ле Гуи – д-р наук (Франция)

И.Н. Лебедев – д-р биол. наук, проф. (Россия)

Л.А. Лутова – д-р биол. наук, проф. (Россия)

Б. Люгтенберг – д-р наук, проф. (Нидерланды)

В.Ю. Макеев – чл.-кор. РАН, д-р физ.-мат. наук (Россия)

В.И. Молодин – академик РАН, д-р ист. наук (Россия)

М.П. Мошкин – д-р биол. наук, проф. (Россия)

С.Р. Мурсалимов – канд. биол. наук (Россия)

Л.Ю. Новикова – д-р с.-х. наук (Россия)

Е.К. Потокина – д-р биол. наук (Россия)

В.П. Пузырев – академик РАН, д-р мед. наук (Россия)

Д.В. Пышный – чл.-кор. РАН, д-р хим. наук (Россия)

И.Б. Rogozin – канд. биол. наук (США)

А.О. Рувинский – д-р биол. наук, проф. (Австралия)

Е.О. Рыкова – д-р биол. наук (Россия)

Е.А. Салина – д-р биол. наук, проф. (Россия)

В.А. Степанов – академик РАН, д-р биол. наук (Россия)

И.А. Тихонович – академик РАН, д-р биол. наук (Россия)

Е.К. Хлесткина – д-р биол. наук, проф. РАН (Россия)

Э.К. Хуснутдинова – д-р биол. наук, проф. (Россия)

М. Чен – д-р биол. наук (Китайская Народная Республика)

Ю.Н. Шавруков – д-р биол. наук (Австралия)

Р.И. Шейко – чл.-кор. НАНБ, д-р с.-х. наук (Беларусь)

С.В. Шестаков – академик РАН, д-р биол. наук (Россия)

Н.К. Янковский – академик РАН, д-р биол. наук (Россия)

Scientific Peer Reviewed Journal

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

*Founded in 1997**Published 8 times annually*

DOI 10.18699/VJGB-22-86

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

Editor-in-Chief

A.V. Kochetov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia

Deputy Editor-in-Chief

N.A. Kolchanov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia

I.N. Leonova, Dr. Sci. (Biology), Russia

N.B. Rubtsov, Professor, Dr. Sci. (Biology), Russia

V.K. Shumny, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia

Executive Secretary

G.V. Orlova, Cand. Sci. (Biology), Russia

Editorial board

O.S. Afanasenko, Full Member of the RAS, Dr. Sci. (Biology), Russia

D.A. Afonnikov, Associate Professor, Cand. Sci. (Biology), Russia

L.I. Aftanas, Full Member of the RAS, Dr. Sci. (Medicine), Russia

E.E. Andronov, Cand. Sci. (Biology), Russia

Yu.S. Aulchenko, Dr. Sci. (Biology), Russia

L.A. Bespalova, Full Member of the RAS, Dr. Sci. (Agricul.), Russia

N.P. Bondar, Cand. Sci. (Biology), Russia

S.A. Borinskaya, Dr. Sci. (Biology), Russia

P.M. Borodin, Professor, Dr. Sci. (Biology), Russia

A. Börner, Dr. Sci., Germany

M. Chen, Dr. Sci. (Biology), People's Republic of China

S.A. Demakov, Dr. Sci. (Biology), Russia

T.A. Gavrilenko, Dr. Sci. (Biology), Russia

I. Grosse, Professor, Dr. Sci., Germany

N.E. Gruntenko, Dr. Sci. (Biology), Russia

S.G. Inge-Vechtomov, Full Member of the RAS, Dr. Sci. (Biology), Russia

E.K. Khlestkina, Professor of the RAS, Dr. Sci. (Biology), Russia

E.K. Khusnutdinova, Professor, Dr. Sci. (Biology), Russia

A.V. Kilchevsky, Corr. Member of the NAS of Belarus, Dr. Sci. (Biology), Belarus

S.V. Kostrov, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia

A.M. Kudryavtsev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

D.M. Larkin, Cand. Sci. (Biology), Great Britain

I.N. Lavrik, Dr. Sci. (Biology), Germany

J. Le Gouis, Dr. Sci., France

I.N. Lebedev, Professor, Dr. Sci. (Biology), Russia

B. Lugtenberg, Professor, Dr. Sci., Netherlands

L.A. Lutova, Professor, Dr. Sci. (Biology), Russia

V.Yu. Makeev, Corr. Member of the RAS, Dr. Sci. (Physics and Mathem.), Russia

V.I. Molodin, Full Member of the RAS, Dr. Sci. (History), Russia

M.P. Moshkin, Professor, Dr. Sci. (Biology), Russia

S.R. Mursalimov, Cand. Sci. (Biology), Russia

L.Yu. Novikova, Dr. Sci. (Agricul.), Russia

E.K. Potokina, Dr. Sci. (Biology), Russia

V.P. Puzyrev, Full Member of the RAS, Dr. Sci. (Medicine), Russia

D.V. Pyshnyi, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia

I.B. Rogozin, Cand. Sci. (Biology), United States

A.O. Ruvinsky, Professor, Dr. Sci. (Biology), Australia

E.Y. Rykova, Dr. Sci. (Biology), Russia

E.A. Salina, Professor, Dr. Sci. (Biology), Russia

Y.N. Shavrukov, Dr. Sci. (Biology), Australia

R.I. Sheiko, Corr. Member of the NAS of Belarus, Dr. Sci. (Agricul.), Belarus

S.V. Shestakov, Full Member of the RAS, Dr. Sci. (Biology), Russia

V.A. Stepanov, Full Member of the RAS, Dr. Sci. (Biology), Russia

I.A. Tikhonovich, Full Member of the RAS, Dr. Sci. (Biology), Russia

A.V. Vasiliev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

M.I. Voevoda, Full Member of the RAS, Dr. Sci. (Medicine), Russia

N.K. Yankovsky, Full Member of the RAS, Dr. Sci. (Biology), Russia

I.K. Zakharov, Professor, Dr. Sci. (Biology), Russia

I.A. Zakharov-Gezekhus, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

719

ОТ РЕДАКТОРА

Н.А. Колчанов, Ю.Г. Матушкин

Системная компьютерная биология

721

ОБЗОР

Молекулярные механизмы детерминации клеток сосудистой системы корня *Arabidopsis thaliana* L.

А.Д. Сидоренко, Н.А. Омелянчук, Е.В. Землянская

733

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Компьютерный анализ особенностей регуляции гиперметилированных маркерных генов гепатокарциномы вирусными белками гепатита С.

Е.А. Антропова, Т.М. Хлебодарова, П.С. Деменков, А.С. Вензель, Н.В. Иванисенко, А.Д. Гавриленко, Т.В. Иванисенко, А.В. Адамовская, П.М. Ревва, И.Н. Лаврик, В.А. Иванисенко

743

ОБЗОР

Рациональная метаболическая инженерия *Corynebacterium glutamicum* для продукции L-валина. М.Е. Шереметьева, К.Э. Ануфриев, Т.М. Хлебодарова, Н.А. Колчанов, А.С. Яненко

758

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Стратификация и слоения в фазовых портретах моделей генных сетей.

В.П. Голубятников, А.А. Акиншин, Н.Б. Аюпова, Л.С. Минушкина

765

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Апробация технологии оценки мимики лиц для изучения динамики функциональных состояний человека в ЭЭГ-парадигме покоя. А.Н. Савостьянов, Е.Г. Вергунов, А.Е. Сапрыгин, Д.А. Лебедкин

773

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Разработка нейронной сети для диагностики риска возникновения депрессии по экспериментальным данным стоп-сигнал парадигмы. М.О. Зеленских, А.Е. Сапрыгин, С.С. Таможников, П.Д. Рудыч, Д.А. Лебедкин, А.Н. Савостьянов

Эволюционная компьютерная биология

780

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Программная система на основе 3D симулятора для моделирования эволюции в популяции организмов, обладающих зрительной системой.

А.П. Девятериков, А.Ю. Пальянов

787

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Фосфолипазы A2 человека:

функциональный и эволюционный анализ. И.И. Турнаев, М.Е. Бочарникова, Д.А. Афонников

798

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Промоторы генов, кодирующих β-амилазу, альбумин и глобулин пищевых растений в сравнении с непищевыми, характеризуются более низкой аффинностью к ТАТА-связывающему белку: *in silico* анализ. О.В. Вишневецкий, И.В. Чадаева, Е.Б. Шарыпова, Б.М. Хандаев, К.А. Золотарева, А.В. Казачек, П.М. Пономаренко, Н.Л. Подколотный, Д.А. Рассказов, А.Г. Богомолов, О.А. Подколотная, Л.К. Савинкова, Е.В. Землянская, М.П. Пономаренко

Компьютерная геномика

806

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

FastContext: инструмент для контекстного анализа последовательностей в данных секвенирования нового поколения (NGS). Э. Весна, В.С. Фишман (на англ. языке)

810

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Сравнительный генетический анализ O-антигенов бактерий семейства Oxalobacteraceae: уникальность или тривиальность? С.Д. Афонникова, А.С. Комиссаров, П.Д. Кучур (на англ. языке)

819

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Контекстные сигналы в митохондриальных микроРНК млекопитающих. О.В. Вишневецкий, П.С. Ворожейкин, И.И. Титов

826

ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ

Свойства малого мира научных организаций определяют динамику публикационной активности в области миРНК. А.Б. Фирсов, И.И. Титов

830

Алфавитный указатель авторов статей, опубликованных в журнале в 2022 г.

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

Systems computational biology

721 REVIEW
Molecular mechanisms of vascular tissue patterning in *Arabidopsis thaliana* L. roots.
A.D. Sidorenko, N.A. Omelyanchuk, E.V. Zemlyanskaya

733 ORIGINAL ARTICLE
Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins.
E.A. Antropova, T.M. Khlebodarova, P.S. Demenkov, A.S. Venzel, N.V. Ivanisenko, A.D. Gavrilenko, T.V. Ivanisenko, A.V. Adamovskaya, P.M. Revva, I.N. Lavrik, V.A. Ivanisenko

743 REVIEW
Rational metabolic engineering of *Corynebacterium glutamicum* to create a producer of L-valine. M.E. Sheremetieva, K.E. Anufriev, T.M. Khlebodarova, N.A. Kolchanov, A.S. Yanenko

758 ORIGINAL ARTICLE
Stratifications and foliations in phase portraits of gene network models.
V.P. Golubyatnikov, A.A. Akinshin, N.B. Ayupova, L.S. Minushkina

765 ORIGINAL ARTICLE
Validation of a face image assessment technology to study the dynamics of human functional states in the EEG resting-state paradigm. A.N. Savostyanov, E.G. Vergunov, A.E. Saprygin, D.A. Lebedkin

773 ORIGINAL ARTICLE
Development of a neural network for diagnosing the risk of depression according to the experimental data of the stop signal paradigm. M.O. Zelenskiy, A.E. Saprygin, S.S. Tamozhnikov, P.D. Rudych, D.A. Lebedkin, A.N. Savostyanov

Evolutionary computational biology

780 ORIGINAL ARTICLE
A software system for modeling evolution in a population of organisms with vision, interacting with each other in 3D simulator.
A.P. Devyaterikov, A.Yu. Palyanov

787 ORIGINAL ARTICLE
Human phospholipases A2: a functional and evolutionary analysis.
I.I. Turnaev, M.E. Bocharnikova, D.A. Afonnikov

798 ORIGINAL ARTICLE
Promoters of genes encoding β -amylase, albumin and globulin in food plants have weaker affinity for TATA-binding protein as compared to non-food plants: *in silico* analysis. O.V. Vishnevsky, I.V. Chadaeva, E.B. Sharypova, B.M. Khandaev, K.A. Zolotareva, A.V. Kazachek, P.M. Ponomarenko, N.L. Podkolodny, D.A. Rasskazov, A.G. Bogomolov, O.A. Podkolodnaya, L.K. Savinkova, E.V. Zemlyanskaya, M.P. Ponomarenko

Computational genomics

806 ORIGINAL ARTICLE
FastContext: A tool for identification of adapters and other sequence patterns in next generation sequencing (NGS) data.
E. Viesná, V. Fishman

810 ORIGINAL ARTICLE
Unique or not unique? Comparative genetic analysis of bacterial O-antigens from the Oxalobacteraceae family. S.D. Afonnikova, A.S. Komissarov, P.D. Kuchur

819 ORIGINAL ARTICLE
The context signals of mitochondrial miRNAs (mitomiRs) of mammals.
O.V. Vishnevsky, P.S. Vorozheykin, I.I. Titov

826 ORIGINAL ARTICLE
Small world of the miRNA science drives its publication dynamics.
A.B. Firsov, I.I. Titov

830 Alphabetical author index for the list of papers published in the journal in 2022



Н.А. Колчанов



Ю.Г. Матушкин

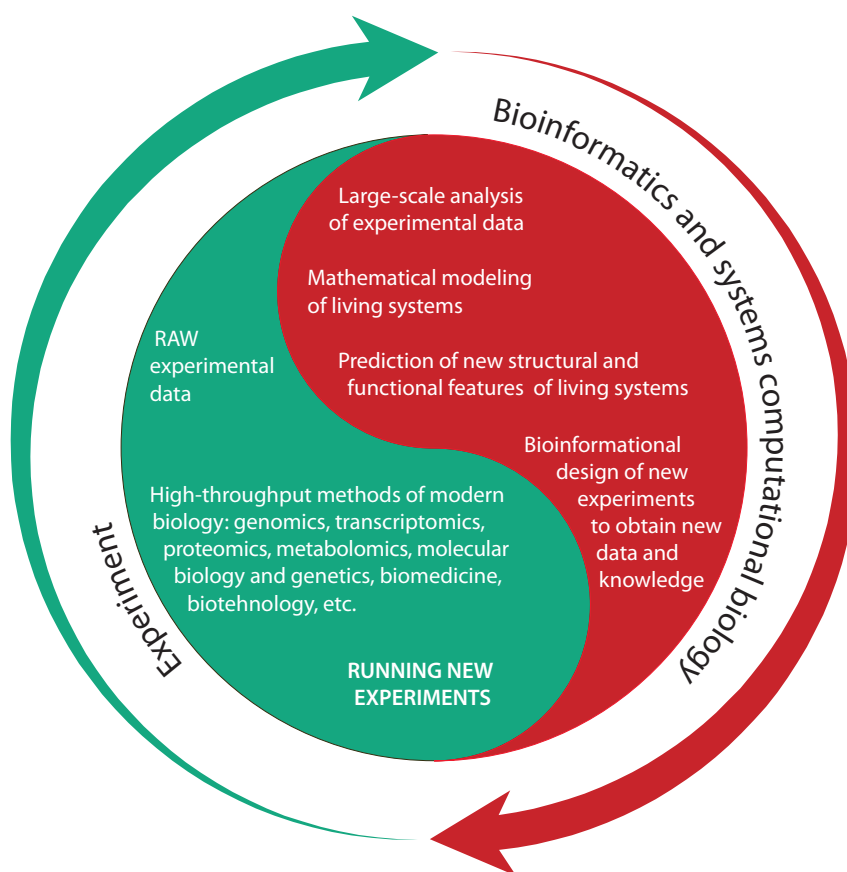
В последние 15 лет в генетике происходит информационный взрыв, обусловленный появлением эффективных методов расшифровки геномов и, как следствие, снижением более чем на четыре порядка стоимости геномного секвенирования. В результате этого науки о жизни стали главным источником больших данных, обогнав по темпам роста не только другие научные направления, но даже социальные сети. Одновременно совершенствовались экспериментальные методы транскриптомного, протеомного, метаболомного и других видов исследований. Все это создало грандиозный вызов для биоинформатики и системной компьютерной биологии, поскольку темпы генерации геномных и других типов данных намного опережают возможности их глубокого компьютерного анализа.

Объем и сложность этих данных настолько велики, что их понимание, интерпретация и, тем более, практическое применение невозможны без использования новых информационных технологий, эффективных методов анализа данных и компьютерного моделирования живых систем. На наших глазах возникает новая биология, ключевой особенностью которой является тесная интеграция экспериментальных и компьютерных подходов. Она включает:

- накопление больших объемов экспериментальных данных о структурной и функциональной организации живых систем, процессов и структур;
- крупномасштабный анализ получаемой экспериментальной информации в контексте накопленных ранее больших данных (молекулярно-биологических, генетических, биохимических, физиологических и др.);
- компьютерное моделирование изучаемых систем, процессов и структур на различных иерархических уровнях их организации;
- предсказание новых свойств и изучаемых живых систем на основе результатов анализа и моделирования;
- планирование новых экспериментов для подтверждения результатов предсказаний и прогнозов и, наконец, проведение новых экспериментов и получение новых данных и знаний.

И вот таким образом стремительно развивается новая биология, движущей силой которой является интеграция экспериментальных и компьютерных подходов. В этой интеграции важнейшую роль играют:

- а) биоинформатика, обеспечивающая хранение, обработку и анализ больших данных, получаемых с помощью методов геномики, транскриптомики, протеомики, метаболомики и других высокопроизводительных экспериментальных технологий, а также
- б) системная компьютерная биология, к числу ключевых задач которой относятся реконструкция, компьютерный анализ и моделирование генных сетей и молекулярно-генетических систем, ответственных за контроль процессов формирования молекулярно-генетических, биохимических, физиологических, структурно-морфологических, поведенческих и других характеристик человека, животных, растений и микроорганизмов на основе информации, закодированной в их геномах.



При этом биоинформатика и системная компьютерная биология имеют важнейшее значение не только для проведения фундаментальных исследований, но и для решения практических задач в интересах сельского хозяйства, биомедицины, фармакологии, биотехнологий, экологии и множества других направлений в науках о жизни и междисциплинарных исследованиях.

В очередной выпуск «Вавиловского журнала генетики и селекции» включены статьи, подготовленные по материалам ряда докладов, представленных на 13-й между-

народной мультikonференции «Биоинформатика регуляции геномов/системная биология» (4–8 июля 2022 г.) по таким направлениям, как компьютерный анализ бактериальных и митохондриальных геномов, компьютерная вирусология (взаимодействия в системе «вирус–хозяин», рациональная инженерия метаболических путей бактерий, моделирование динамики генных сетей и процессы морфогенеза растений, эволюционная компьютерная биология, экспериментально-компьютерная психология и ряд других.

Научные редакторы выпуска:

*академик Н.А. Колчанов,
научный руководитель ФИЦ ИЦиГ СО РАН*

*канд. биол. наук Ю.Г. Матушкин,
вед. науч. сотрудник ФИЦ ИЦиГ СО РАН*


Original Russian text <https://sites.icgbio.ru/vogis/>

Molecular mechanisms of vascular tissue patterning in *Arabidopsis thaliana* L. roots

A.D. Sidorenko^{1, 2}, N.A. Omelyanchuk¹, E.V. Zemlyanskaya^{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

 ezemlyanskaya@bionet.nsc.ru

Abstract. A vascular system in plants is a product of aromorphosis that enabled them to colonize land because it delivers water, mineral and organic compounds to plant organs and provides effective communications between organs and mechanical support. Vascular system development is a common object of fundamental research in plant development biology. In the model plant *Arabidopsis thaliana*, early stages of vascular tissue formation in the root are a bright example of the self-organization of a bisymmetric (having two planes of symmetry) pattern of hormone distribution, which determines vascular cell fates. In the root, vascular tissue development comprises four stages: (1) specification of progenitor cells for the provascular meristem in early embryonic stages, (2) the growth and patterning of the embryo provascular meristem, (3) postembryonic maintenance of the cell identity in the vascular tissue initials within the root apical meristem, and (4) differentiation of their descendants. Although the anatomical details of *A. thaliana* root vasculature development have long been known and described in detail, our knowledge of the underlying molecular and genetic mechanisms remains limited. In recent years, several important advances have been made, shedding light on the regulation of the earliest events in provascular cells specification. In this review, we summarize the latest data on the molecular and genetic mechanisms of vascular tissue patterning in *A. thaliana* root. The first part of the review describes the root vasculature ontogeny, and the second reconstructs the sequence of regulatory events that underlie this histogenesis and determine the development of the progenitors of the vascular initials in the embryo and organization of vascular initials in the seedling root.

Key words: meristem; xylem; phloem; (pro)cambium; plant hormones; auxin; cytokinin; *Arabidopsis thaliana*.

For citation: Sidorenko A.D., Omelyanchuk N.A., Zemlyanskaya E.V. Molecular mechanisms of vascular tissue patterning in *Arabidopsis thaliana* L. roots. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):721-732. DOI 10.18699/VJGB-22-88

Молекулярные механизмы детерминации клеток сосудистой системы корня *Arabidopsis thaliana* L.

А.Д. Сидоренко^{1, 2}, Н.А. Омелянчук¹, Е.В. Землянская^{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

 ezemlyanskaya@bionet.nsc.ru

Аннотация. Сосудистая система является результатом ароморфоза, который позволил растениям успешно освоить сушу. За счет нее осуществляется проведение воды, минеральных и органических соединений, обеспечивается эффективное сообщение между органами, а также выполняется функция механической опоры. Процесс формирования сосудистой системы – общепринятый объект фундаментальных исследований в области биологии развития растений. В частности, ранние этапы развития сосудистой системы корня модельного растения *Arabidopsis thaliana* представляют собой яркий пример самоорганизации бисимметричного (имеющего две плоскости симметрии) паттерна распределения фитогормонов, который направляет детерминацию клеток сосудистой системы. В процессе формирования сосудистой системы корня можно выделить четыре этапа: 1) детерминацию (спецификацию) клеток-предшественников проваскулярной меристемы на ранних стадиях эмбриогенеза; 2) рост и разметку проваскулярной меристемы зародыша; 3) постэмбриональное поддержание инициалей (стволовых клеток) сосудистой системы в апикальной меристеме корня; 4) конечную специализацию (дифференцировку) их дочерних клеток. Анатомические детали развития сосудистой системы *A. thaliana* давно известны и подробно описаны, однако наши знания о молекулярно-генетических механизмах этого процесса все еще ограничены. В последние годы сделано несколько важных открытий, проливающих свет на регуляцию самых ранних событий, предшествующих дифференцировке клеток сосудистой системы. В настоящем обзоре мы обобщаем данные о молекулярно-генетических механизмах, определяющих

направление клеточной дифференцировки в элементы сосудистой системы корня у *A. thaliana*. Первая часть обзора посвящена описанию гистогенеза сосудистой системы корня. Далее мы реконструируем последовательность регуляторных событий, которые лежат в основе этого гистогенеза и обуславливают развитие предшественников инициалей сосудистой системы у зародыша и организацию инициалей сосудистой системы в корне проростка.

Ключевые слова: меристема; ксилема; флоэма; (про)камбий; фитогормоны; ауксин; цитокинин; *Arabidopsis thaliana*.

Introduction

Evolutionary formation of a vascular system in plants was a necessary prerequisite for terrestrial colonization (Lucas et al., 2013). Vasculature provides mechanical support, effective transportation of water, and mineral and organic compounds as well as signal molecules and by this has enabled plants to reach enormous sizes and populate different territories. The vascular system consists of two domains different in their structure and functions. These are xylem that provides water transportation and delivers mineral compounds from the root to above-ground organs; and phloem that conveys organic compounds from photosynthesizing tissues rootward (Evert, Eichhorn, 2006).

In angiosperms, the mature xylem consists of (1) water-transportation vessels; (2) fibers to provide mechanical support; (3) parenchyma cells (Evert, Eichhorn, 2006). The vessels are the hollow tubes formed by the cells connected in a row and having perforations in the *anticlinal* walls and pores in the *periclinal* walls (Fig. 1). The vessels and fibers are a product of the programmed death of the cells that have formed a ligni-

fied secondary cell wall (Courtois-Moreau et al., 2009; Smith et al., 2013; Furuta et al., 2014). Meanwhile, the living cells of parenchyma perform a storage function, participating in vessel lignification and regulating the water transport speed (Ménard, Pesquet, 2015; Růžicka et al., 2015).

The phloem, on the other hand, consists of (1) sieve tubes to transport organic substances; (2) companion cells; (3) fibers and sclereids to provide mechanical support, and (4) parenchyma cells (Sjolund, 1997; Evert, Eichhorn, 2006). Unlike the lignified hollow vessels of the xylem, the sieve tubes are a strand of living cells (sieve elements) communicating by sieve fields, anticlinal-wall regions with high numbers of small pores. The sieve elements form a thickened non-lignified secondary cell wall (Heo et al., 2014) and their main feature is the lack most of the organelles including a nucleus, vacuole, rough endoplasmic reticulum, Golgi body, cytoskeleton, ribosomes whose presence could prevent substances transportation. The viability of the sieve elements is maintained by companion cells – the parenchyma cells with large nuclei and mitochondria, directly contacting sieve elements. As for

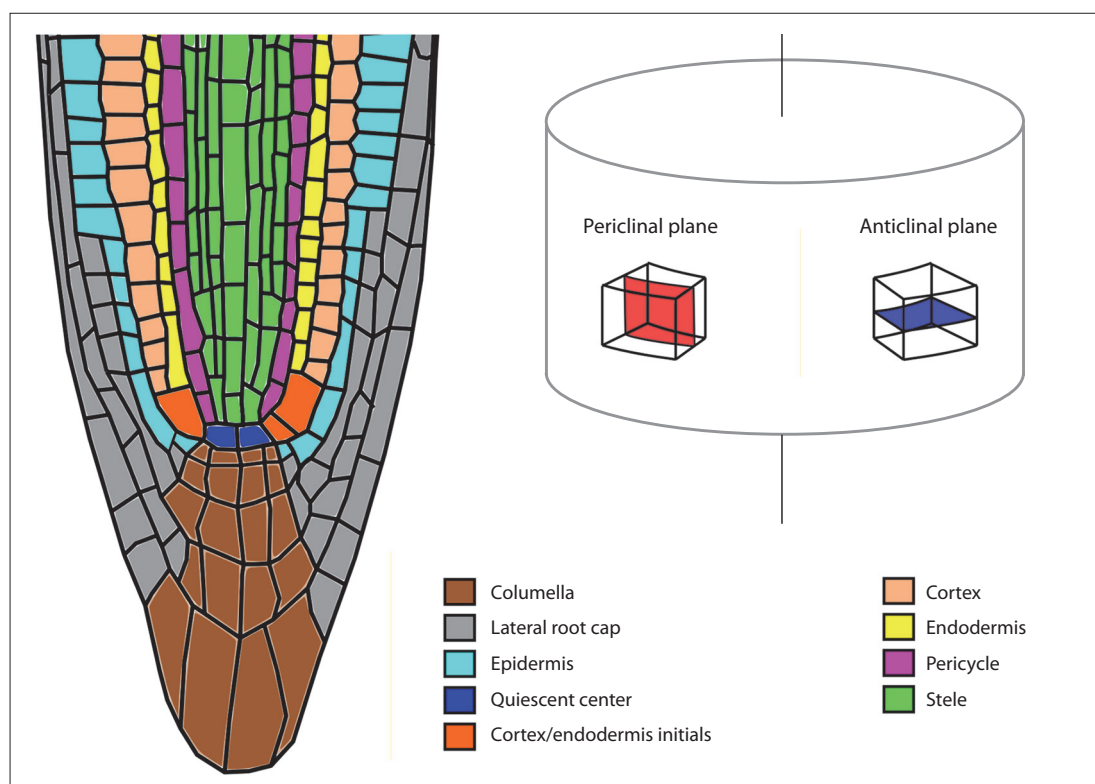


Fig. 1. *Arabidopsis thaliana* root apical meristem.

The vertical line is the root central axis.

mechanical phloem elements – fibers and sclereids – they differ from each other by the shape of their cells. While the former are strongly elongated and pointed at the ends, the latter are just slightly elongated.

Organization of vascular system is different for different organs in different plant species at different stages of their development (Scarpella, Meijer, 2004; Lucas et al., 2013; Furuta et al., 2014). Nevertheless, the mechanisms determining its development are quite conservative (Li et al., 2010; Seo et al., 2020). Plant cells are not capable of migration, so during morphogenesis, the tissue and organ architecture is formed by regulating the sequence and orientation of cell divisions. In terms of its anatomy, the vascular system development has been described in much detail (Scheres et al., 1994; Evert, Eichhorn, 2006; Miyashima et al., 2013; Furuta et al., 2014; De Rybel et al., 2014b, 2016), however, the molecular and genetic mechanisms responsible for this process are much less known. Our current understanding of these mechanisms is mainly based on the investigation of the model plant *Arabidopsis thaliana*.

In the further sections of this review, we will provide a short description of vascular tissue histogenesis in this plant species and reconstruct the corresponding sequence of regulatory events. We will describe the control of root vascular system development in the embryo and seedling, i.e. the earliest stages of its formation. As for the mechanisms controlling vasculature development at later stages, their description can be found in the recent reviews (see Agustí, Blázquez, 2020; Seo et al., 2020).

Root vascular tissue histogenesis

There are primary (produced by the *primary meristem*) and secondary (produced by the *secondary meristem*) vascular tissues.

Development of root primary conductive tissues

At the globule stage of *A. thaliana* embryogenesis the specification of four *provascular initials* occurs. Provascular initials undergo oriented divisions, finally giving rise to the *provascular meristem* of the embryonic root and hypocotyl (Fig. 2) (Scheres et al., 1994; Evert, Eichhorn, 2006; Miyashima et al., 2013; Furuta et al., 2014; De Rybel et al., 2014b, 2016). The cells of provascular meristem are not yet differentiated, but the cellular fate of some of them has already been determined – after the embryo germination they give birth either to xylem or to phloem cells. The positions of these predetermined cells in the provascular meristem matches that of the bisymmetric (that is, having two planes of symmetry) *diarch* organization of the vascular system in the postembryonic root tip: in its transverse – section, there is one layer of xylem precursor cells surrounded on both sides by *procambial* cells that separate the future xylem from two files of phloem progenitor cells, which lie in a perpendicular plane (Dolan et al., 1993) (Fig. 2, 3, a). This structure is surrounded by *pericycle* cells that are also derived from provascular initials, so together they form a *central cylinder* or a *stele* (see Fig. 1). It is noteworthy that the terminology designating the cells in developing root vascular system is rather blurred (Furuta et al., 2014). In particular, the term ‘procambium’ is applied to address either indeterminate

Glossary

Amphicribal vascular bundle – a vascular bundle in which the phloem surrounds the xylem.

Anticlinal – located in a plane perpendicular to the surface of a tissue or organ. Talking about anticlinal cell walls or divisions we will mean an anticlinal plane perpendicular to the central axis of an organ.

Anticlinal cell division – cell division in the anticlinal plane that leads to an increase in length.

Asymmetric cell division – results in the formation of two daughter cells with different cell fates.

Cortex – a cell layer surrounding the endodermis.

Diarch vascular bundle – a vascular bundle whose phloem and xylem are located at different radii, wherein two rays of xylem are distinguished.

Endodermis – the innermost cell layer surrounding the stele.

Hypophysis is the upper cell of the suspensor, which acquires its identity at the 16–32 cell stage; gives rise to the quiescent center (the organizing center of the root apical meristem) and the root cap.

Periclinal – located in a plane parallel to the surface of a tissue or organ.

Periclinal cell division – cell division in the periclinal plane leading to an increase in the number of cell layers in the radial direction.

Pericycle – parenchyma cell layer surrounding conductive tissues and forming the stele outer layer.

Primary meristem – formed during embryogenesis.

Procambium – indeterminate primary vascular meristem cells located between the xylem plate and phloem poles in the root of *Arabidopsis thaliana*.

Provascular initials – four proembryo cells occurring at the early globular stage to form the entire provascular meristem of the root/hypocotyl, and only it.

Provascular root/hypocotyl meristem – primary meristem from which the primary vascular system of these organs differentiates after embryo germination.

Root apical meristem – primary root meristem to produce all cells of the root during its post-embryonic growth.

Secondary meristem – formed during the postembryonic period.

Stele (central cylinder) – primary conductive tissues located in the center of the axial organ, and surrounded by a pericycle.

Suspensor – a structure at the base of an embryo that connects it to endosperm and consists of the descendants of a two-celled pro-embryo basal cell.

Vascular cambium – secondary vascular meristem to provide root thickening.

Xylem plate – a layer of primary xylem cells (or their predetermined precursors) located in the central plane along the root axis.

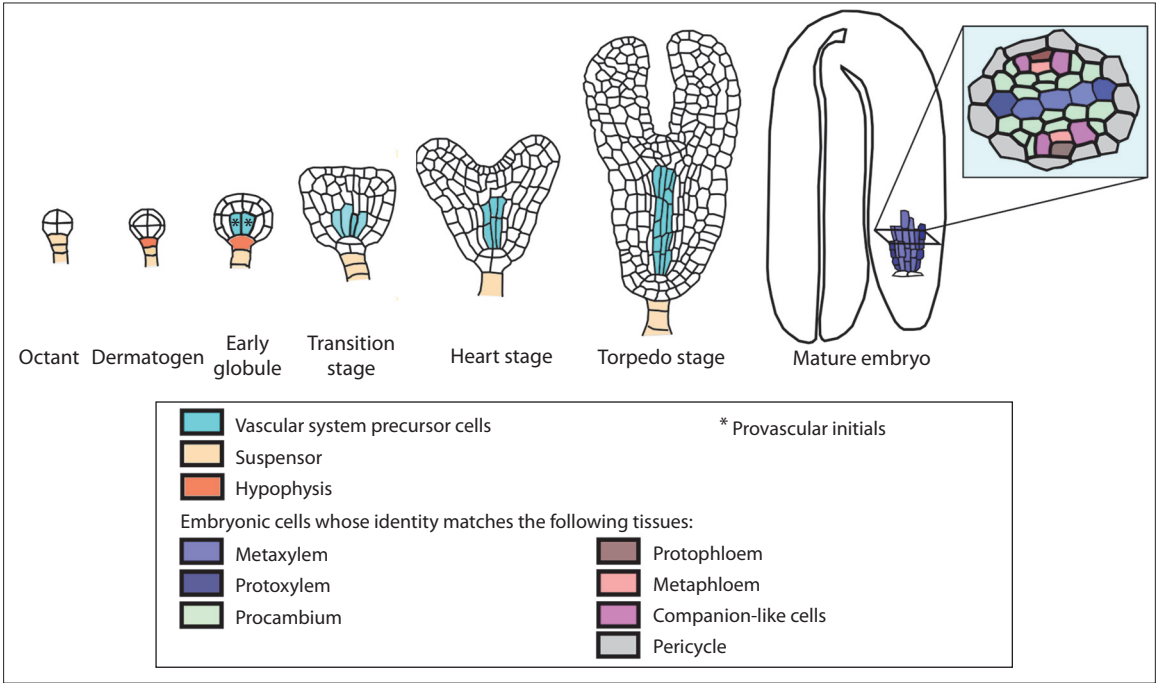


Fig. 2. Provascular meristem development in *A. thaliana* embryo.
The mature embryo contains predetermined but not differentiated progenitor cells of the future vascular system elements.

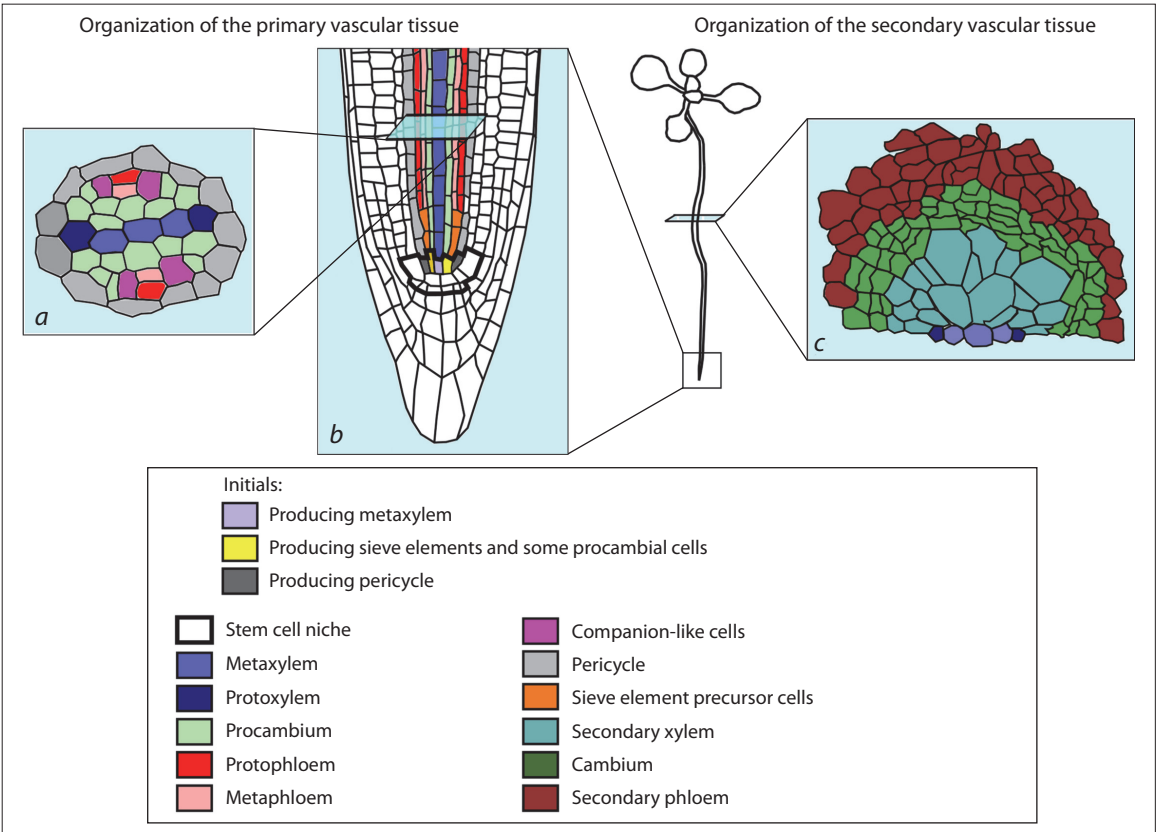


Fig. 3. Primary and secondary vascular tissues in postembryonic *A. thaliana* root.
The root-tip stele (*a*, *b*) is diarch and comprised of the procambium, primary phloem and xylem surrounded by pericycle. The primary phloem is composed of proto- and metaphloem and of companion-like cells. The primary xylem consists of proto- and metaxylem. In the stem cell niche in longitudinal section (*b*) two initials producing the procambium/proto-/metaphloem, two initials producing the pericycle, and one producing the metaxylem are visible. During the root secondary growth (*c*), the cambium produces phloem cells outwards and xylem cells – inwards, so the vascular bundle stops being diarch and becomes amphicribal.

cells of the primary vascular tissue in seedlings (and their progenitors) or the whole embryonic provascular meristem (see Busse, Evert, 1999).

Soon after germination, vascular elements start to differentiate in a hypocotyl stele and cotyledon veins, the provascular meristem of the latter comes from the shoot apical meristem (Miyashima et al., 2013). From the hypocotyl, the process spreads upwards and downwards taking the epicotyl and root, respectively (Busse, Evert, 1999; De Rybel et al., 2014b; Furuta et al., 2014). In *A. thaliana*, these are the protophloem sieve elements adjacent to the pericycle that differentiate first, and since the cells surrounding them keep elongating, protophloem cells soon die to be functionally replaced by the metaphloem sieve elements placed closer to the center of the stele (Graeff, Hardtke, 2021; Truernit, 2022). Later, the protoxylem vascular elements are formed that are located at the poles of the **xylem plate** and have annular or spiral thickenings of the secondary cell walls. The last cells to differentiate are the metaxylem cells occupying the central position in the xylem plate and having pitted or reticulate lignin deposits (Růžicka et al., 2015).

While the root grows in length, its new cells are produced through **anticlinal division** of the cells in the **apical meristem** located at the root tip (Desvoyes et al., 2021). In *A. thaliana*, the root apical meristem is closed, i.e. different stem cells (initials) can produce not any but strictly limited set of cell types and for each differentiated cell it is easy to trace which stem cell it has originated from (see Fig. 1). Among the stele initials those can be distinguished that give birth to (1) protoxylem; (2) metaxylem; (3) procambium and sieve elements of proto- and metaphloem (in this case, the three cell types are produced through a series of anticlinal and **periclinal divisions**); (4) only procambial cells; (5) pericycle (see Fig. 3, b) (Mähönen et al., 2000; Rodriguez-Villalon et al., 2015; Truernit, 2022). The mutual arrangement of initials corresponds to the diarch organization of young root vasculature, so the cell identity established in the embryo provascular meristem is maintained in the root apical meristem. Here it is worth mentioning that apart from the proto- and metaxylem, proto- and metaphloem and procambium there are also companion cells. Some authors designate them more strictly as companion-like cells (Truernit, 2022). These cells are adjacent to the sieve elements of proto- and metaphloem and possess a number of morphological and physiological characteristics of companion cells (Stadler et al., 2005; Ross-Elliott et al., 2017; Smetana et al., 2019; Graeff, Hardtke, 2021) but, unlike the latter, they do not share a common initial with the proto- and metaphloem elements in the stem cell niche (Mähönen et al., 2000). The companion-like cells differentiate when the protophloem sieve elements start functioning (Graeff, Hardtke, 2021). In *A. thaliana*, the xylem and phloem parenchyma, fibers and true companion cells differentiate only during the secondary growth (Růžicka et al., 2015; Truernit, 2022).

Cambium formation

In *A. thaliana* primary vascular system, periclinal divisions of procambium cells are few, but after differentiation of the primary vascular elements, these cells begin to actively divide periclinally. The periclinal divisions also occur in the pericycle

cells adjacent to the xylem plate. As a result, a closed cell ring forms around the xylem to give birth to the **vascular cambium** (see Fig. 3, c) (Baum et al., 2002; Nieminen et al., 2015; Růžicka et al., 2015; Smetana et al., 2019). It is noteworthy that only those procambium and pericycle cells in direct contact with the xylem primary vessels give rise to the vascular cambium, i.e., have the properties of stem cells (Smetana et al., 2019) while the descendants of other proliferating procambial cells differentiate into the phloem.

Thus, the diarch root vasculature transforms into **amphicribal** one, in which the xylem is surrounded by the phloem with the cambium placed in between (see Fig. 3, c). Through **asymmetric division**, every initial is capable of producing phloem cells outwards and xylem cells inwards, so the root gets thicker (Smetana et al., 2019). In some species, e.g., in the vast majority of monocots, the cambium is not formed and no secondary growth is initiated. In this case, all procambium cells get differentiated.

Embryo polarity establishment and the predetermination of provascular initials

The development of a multicellular organism is accompanied by a gradual increase in the limitation of cellular potencies. At the first stage of this process predetermination or specification occurs, in other words, the fate of a totipotent cell is established in terms of the progenitor of what type of cells it will become. Meanwhile, the cell remains undifferentiated and can change its fate under certain conditions. The process of cell identity determination involves the local accumulation of signal molecules, which either activate or suppress the activity the gene networks inherent in specific cell types. In this case, an important role is given to the non-cell-autonomous factors able to move between cells and form gradients (Seo et al., 2020).

Provascular stem cells specification at the early globular stage of embryogenesis is preceded by a series of cell divisions and embryo polarity determination (Lau et al., 2012; De Rybel et al., 2014b). The proper accomplishment of these processes is essential for the vascular tissue to begin its development from the right number of cells placed in the right positions. Plant hormone auxin is a key regulator of embryogenesis, whose heterogeneous distribution provides positional information, which directs embryo development (Weijers, Jürgens, 2005; Smit, Weijers, 2015; Mironova et al., 2017). The main auxin effector in embryogenesis is transcription factor (TF) AUXIN RESPONSE FACTOR 5 (ARF5)/MONOPTEROS (MP) (Smit, Weijers, 2015; Verma et al., 2021) and it is believed that forming the auxin signal-distribution pattern is provided mainly due to feedbacks in regulation of phytohormone biosynthesis, its polar intercellular transport and signaling pathway (Sauer et al., 2006; Möller, Weijers, 2009; Lau et al., 2011; Robert et al., 2015). As a result, at the early stages of embryogenesis, auxin is accumulated in the apical cells to determine the embryo polarity (Wabnik et al., 2013). Starting from the early globular stage (32 cells), its maximum is shifted to the upper cells of the **suspensor** including the **hypophysis** that later gives rise the quiescent center of the root apical meristem (Friml et al., 2003; Tanaka et al., 2006).

Although the four provascular initials are only distinguished at the early globular stage (Scheres et al., 1994), the cellular

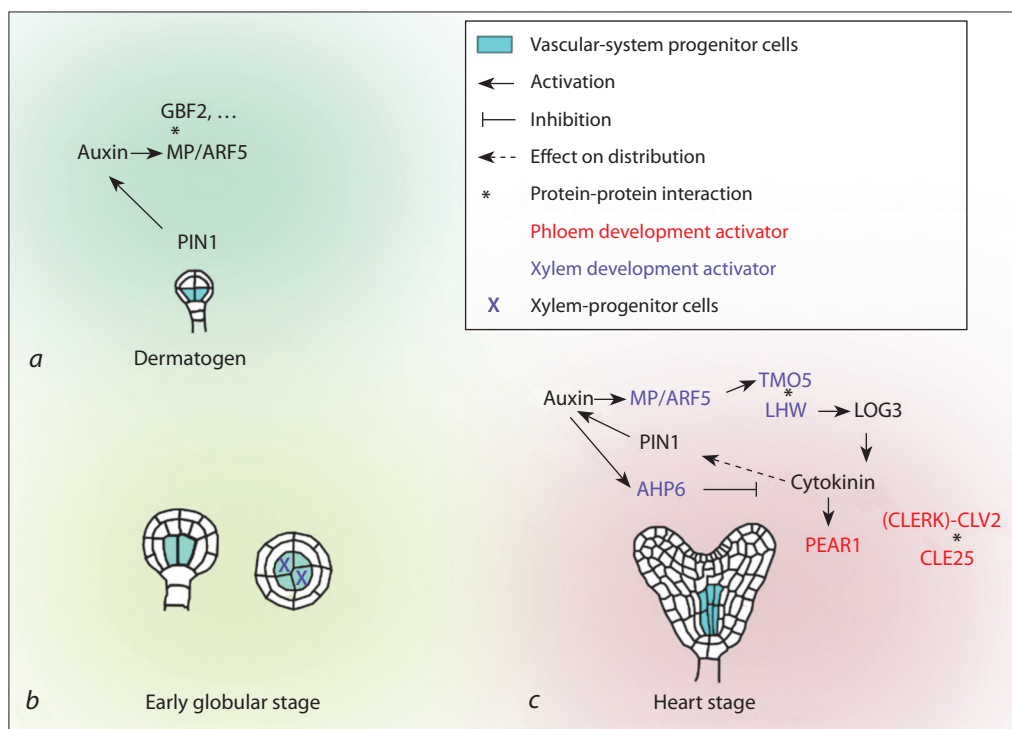


Fig. 4. Genetic regulation of provascular meristem development during embryogenesis.

a, Predetermination of provascular initials. The identity of vascular-tissue progenitors is determined in the four inner cells of the proembryo lower layer at the dermatogen stage. However, anatomically the four initials can only be detected at the early globular stage. **b**, Xylem progenitor predetermination at the early globular stage; **c**, formation of the bisymmetric pattern and xylem/phloem progenitor predetermination starting from the heart stage.

identity of vascular tissue progenitors is determined in the four inner cells of the lower layer of the proembryo as early as at dermatogen stage (Fig. 4, *a*) (Smit et al., 2020). Via periclinal division at transferring to the 32-cell stage, they produce outwards the **ground tissue** progenitors that lose the vascular identity of their maternal cells (see Fig. 4, *a, b*) (Palovaara et al., 2017; Smit et al., 2020). A necessary condition for provascular-initial specification is ARF5/MP-dependent activation of the auxin signaling pathway, but meeting this condition alone is not enough (Möller et al., 2017; Smit et al., 2020). While particular auxin assistants remain unknown, it is suggested that this role is performed not by a single key regulator but by a multicomponent regulatory network, and TF G-BOX BINDING FACTOR 2 (GBF2) is believed to be one of its members (Smit et al., 2020) (see Fig. 4, *a*). GBF2 is assumed to modulate ARF5/MP binding to target-gene promoters. It is worth mentioning here that the state, in which vascular system progenitors are uniformly specified is most likely transient with no stable uniform cellular identity.

Vascular cell predetermination in the provascular meristem

As the oriented divisions of the provascular initials and their descendants continue, the hypocotyl and root vascular systems become patterned through specification of particular cellular types. An important aspect at this stage is setting the boundaries for the cellular domains with different structural and functional identities. By the end of embryogenesis, in

the embryo provascular meristem, the cell identity of all elements such as proto- and metaphloem, proto- and metaxylem, companion-like cells and procambium has been determined as evidenced by the data on cell morphology and expression of marker genes (see Fig. 2) (Bonke et al., 2003; Bauby et al., 2007).

In *A. thaliana*, the bisymmetry of the future root is believed to be predetermined already at the early globular stage by the extended contact between two provascular initials located diagonally relative to each other (see Fig. 4, *b*). This contact is probably formed due to the inaccurate match of cell division planes in proembryo (at the four-cell stage) and is important for xylem plate formation (De Rybel et al., 2014a). Starting from the early heart stage, auxin begins to be actively transported into such contacting provascular cells from the cotyledon primordia located above them, while in other cells the hormone levels remain low (Bishopp et al., 2011a; Help et al., 2011; De Rybel et al., 2014a). The local increase in auxin concentration is necessary for the specification of xylem progenitor cells (Bishopp et al., 2011a).

At the same time, the cells rich in auxin begin to act as an organizing center for the provascular meristem, coordinating its growth through periclinal divisions and establishment of bisymmetric organization (De Rybel et al., 2014a). Auxin induces the ARF5/MP-dependent expression of TFs TARGET OF MONOPTEROS 5 (TMO5) and TMO5-LIKE1 (TSL1) (Schlereth et al., 2010; De Rybel et al., 2013, 2014b), which, forming heterodimers with the auxin-independent LONE-

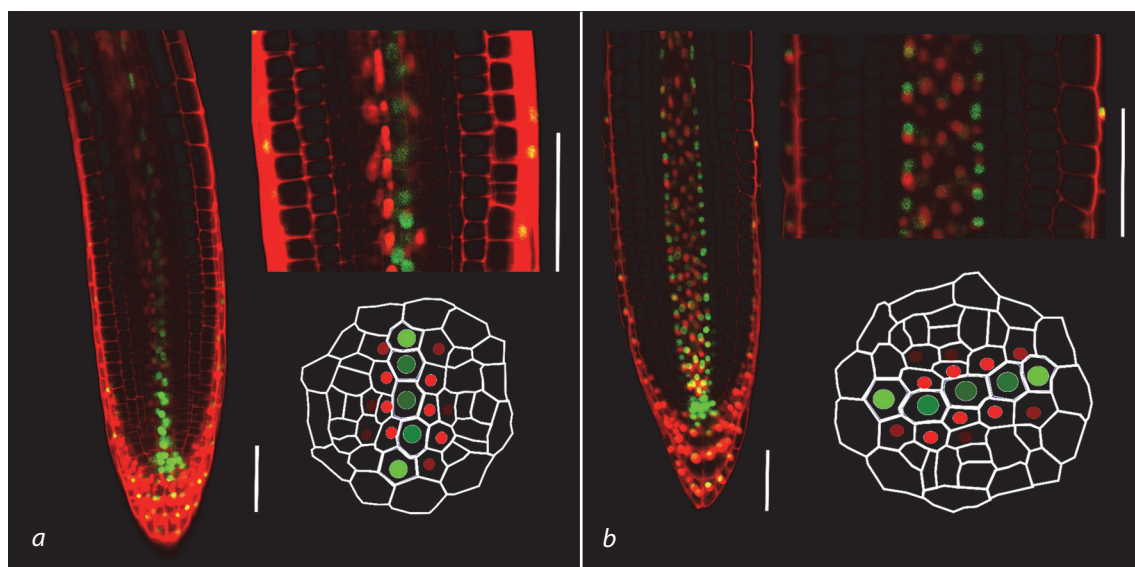


Fig. 5. Bisymmetric auxin/cytokinin distribution pattern in *A. thaliana* root tip stele. *a*, Xylem plate is located perpendicular to an optical section plane; *b*, xylem plate is in parallel to an optical section plane.

Microimages for the *TCSn::ntdTomato-DR5revV2::n3GFP* reporter line (Smet et al., 2019) were obtained using a confocal microscope. The cell walls were stained with propidium iodide. GFP (green) and Tomato (red) nuclear signals mark the activity of auxin and cytokinin signaling pathways, respectively. An auxin response is observed in xylem progenitors with the maximum in protoxylem ones, and a cytokinin response – in xylem-adjacent procambial cells, in this way marking the morphofunctional domains of the root tip stele. Scale 50 μ m.

SOME HIGHWAY (LHW) TF (De Rybel et al., 2013), activate the expression of cytokinin biosynthesis genes *LONELY GUY3* (*LOG3*) and *LOG4* (Kuroha et al., 2009; De Rybel et al., 2014a) (see Fig. 4, *c*). Simultaneously, auxin blocks cytokinin signal transduction, increasing the expression of gene *ARABIDOPSIS HISTIDINE PHOSPHOTRANSFER PROTEIN 6* (*AHP6*) encoding a cytokinin signaling pathway inhibitor (Mähönen et al., 2006; Bishopp et al., 2011a), so a local cytokinin source is formed in xylem progenitors lacking cytokinin signaling.

The high cytokinin level, on the one hand, limits auxin efflux from xylem progenitor cells by controlling the localization of auxin transporter PIN-FORMED 1 (*PIN1*) on the cell membrane (Marhavý et al., 2011; De Rybel et al., 2014a). On the other hand, cytokinin diffuses into neighboring cells following the concentration gradient. In these cells, in the absence of the inhibitor (Cheng, Kieber, 2014), cytokinin activates signaling cascade to stimulate periclinal divisions (Smit, Weijers, 2015). Simultaneously, cytokinin signaling suppresses cell specification into xylem (Mähönen et al., 2006). This mechanism provides for the radial growth of the provascular meristem, which is accompanied by spatial separation of the domains for increased auxin signal (cells obtain xylem identity) and cytokinin signal (pluripotent procambial cells). Its sufficiency for self-organization of the bisymmetric pattern was confirmed using a mathematical model (De Rybel et al., 2014a).

In early embryogenesis, provascular-meristem progenitors begin to express genes encoding peptide hormone *CLAVATA 3* (*CLV3*)/*EMBRYO SURROUNDING REGION 25* (*CLE25*) and mobile TFs of the DNA BINDING WITH ONE FINGER (DOF) family united in the *PHLOEM EARLY DOF* (*PEAR*) group marking sieve-element progenitors in the postembryonic period (Miyashima et al., 2019; Ren et al., 2019). *CLE25*

is expressed starting from a 64-cell embryo stage (Ren et al., 2019). Cytokinin-independent expression of *PEAR1* is detected already at a 16-cell stage, and starting from an early heart stage, this gene expression is activated by cytokinin (Miyashima et al., 2019). It is assumed that the *CLE25* peptide binding to the *CLE-RESISTANT RECEPTOR KINASE* (*CLERK*)-*CLV2* receptor together with the *PEAR1* TF contribute to the early specification of phloem progenitor cells. However, unlike that for xylem, the mechanism to initiate phloem development in embryogenesis remains unknown.

Maintaining xylem/phloem-precursor cellular identity in the root apical meristem

Bisymmetric pattern in stele

In the postembryonic period, the stele cells progenitors maintain the bisymmetric pattern established in embryogenesis, so some of the mechanisms regulating the cell dynamics and vascular-system element predetermination in provascular meristem keep functioning even after germination. However, it cannot be said with complete certainty that these mechanisms are identical.

In the apical meristem, auxin-rich xylem progenitors retain the function of an organizing center, carrying out TMO5/LHW-mediated regulation of cytokinin levels in procambial cells (Fig. 5) (Ohashi-Ito, Bergmann, 2007; Bishopp et al., 2011a; De Rybel et al., 2013; Ohashi-Ito et al., 2013, 2014; Vera-Sirera et al., 2015; Yang et al., 2021). The high content of active cytokinin in xylem cells is maintained by TMO5/LHW-dependent activation of not only cytokinin biosynthesis genes *LOG3* and *LOG4* but also of the *BGLU44* gene encoding a β -glucosidase enzyme (Fig. 6). Cytokinin response in xylem is blocked by auxin through *AHP6* gene expression

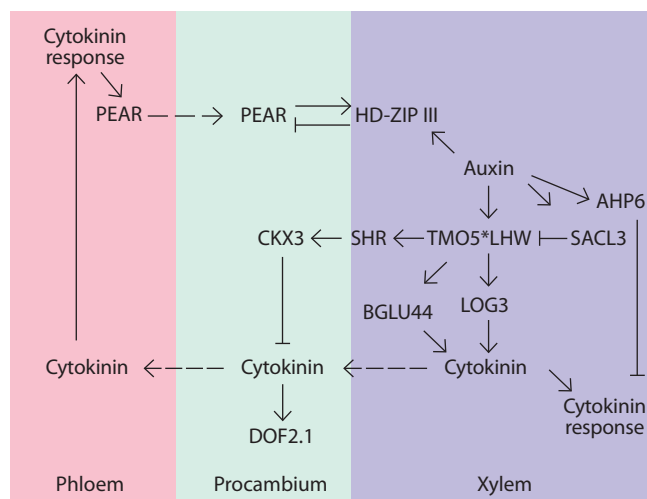


Fig. 6. Maintaining the bisymmetric pattern of auxin/cytokinin distribution in *A. thaliana* root tip stele during the postembryonic period.

The star marks a physical interaction between proteins to form a dimer. The dashed arrow indicates mobile regulator movement.

induction (Bishopp et al., 2011a) as well as through limiting the activity of TMO5/LHW by activating the ACAULIS 5 (ACL5)–SUPPRESSOR OF ACAULIS5 LIKE3 (SACL3) regulatory module blocking the formation of the TMO5/LHW heterodimer by competing with TMO5 for binding to LHW (Katayama et al., 2015; Cai et al., 2016) (see Fig. 6). Meanwhile, in xylem-adjacent procambial cells, the level of the cytokinin diffusing from the xylem is limited by TMO5/LHW-dependent activation of *CYTOKININ OXIDASE 3* (*CKX3*). The activation is mediated by the mobile *SHORT ROOT* (*SHR*) TF, encoded by *TMO5/LHW* target gene. The combined action of multidirectional regulatory modules ensures the stability of the pattern to short-term fluctuations in auxin concentrations in xylem cells, while maintaining its sensitivity to slower/stable changes (Yang et al., 2021). What is interesting is that the *SHR* gene is important not only for the root radial symmetry but also for the functioning of the quiescent center (Tvorogova et al., 2012).

TMO5/LHW-induced cytokinin activates the transcription of the *DOF2.1* TF in the procambial cells surrounding the xylem pole, thus controlling their division (see Fig. 6) (Smet et al., 2019). It is worth noting that, besides xylem cells, it is differentiated phloem that transports the phytohormone and thus can be a source of cytokinin in the root apical meristem (Bishopp et al., 2011b). However, mathematical modeling has demonstrated that phloem cytokinin is not a fundamental source of the positional information for bisymmetric pattern formation (Muraro et al., 2014). At the same time, the high cytokinin content at the phloem poles arranges periclinal divisions of procambium cells through activating the genes of mobile TFs of the *DOF* family united in the *PEAR* group including *PEAR1*, *PEAR2*, *TMO6*, *DOF6* (Miyashima et al., 2019; Smet et al., 2019). They create a concentration gradient and activate the periclinal divisions of the procambial cells surrounding the phloem pole. *HOMEODOMAIN LEU-ZIPPER* class-III (*HD-ZIP III*), TFs whose expression domain is set in

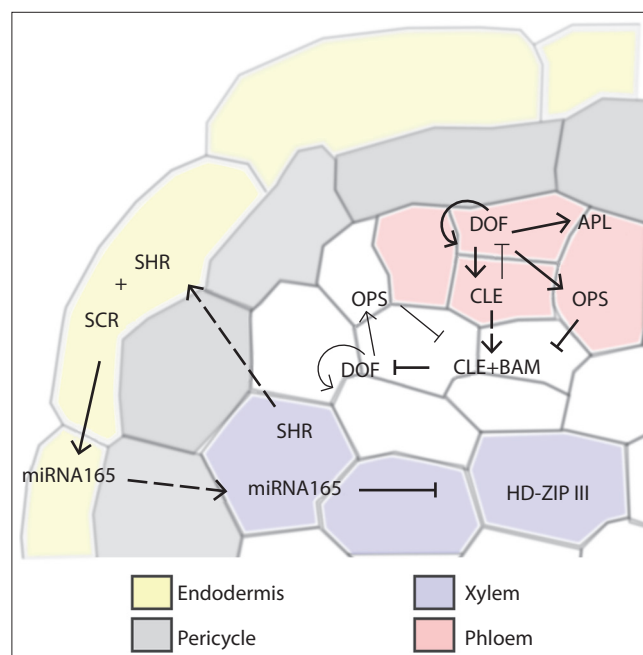


Fig. 7. Genetic circuit regulating proto- and metaxylem/phloem cell predetermination in *A. thaliana*.

Separation of the proto- and metaxylem domains is determined by the concentration gradient of the TFs of the *HD-ZIP III* TF family. The auxin-activated mobile *SHR* TF diffuses from the xylem to the endodermis and binds to the *SCR* protein to activate *miRNA165* expression. MicroRNAs that degrade *HD-ZIP III* family TF mRNA form the concentration gradient towards the center and limit *HD-ZIP III* TF localization to the central domain, thus predetermining metaxylem cells. Phloem predetermination, on the other hand, begins with cytokinin-activated expression of the *DOF* family TFs. They activate the signal *CLE* peptides that migrate to neighboring cells, interact with the *BAM* receptors, and induce *DOF* degradation to produce a boundary between the future phloem and its neighboring cells. The dashed arrow indicates mobile regulator movement.

the central part of the stele (see below) limit the activity of the *PEAR* TFs (see Fig. 6), and *PEAR1* activates the transcription of the genes belonging to the *HD-ZIP III* family, forming a negative feedback loop.

Proto- and metaxylem predetermination

As in embryogenesis, auxin is necessary for xylem cells predetermination in the root apical meristem. In proto- and metaxylem predetermination, a key role is given to the *SHR* and *miRNA165/166* mobile regulators (Fig. 7). *SHR* is produced by xylem cells, from where the TF spreads towards the periphery and, upon reaching the *endodermis*, activates the *SCARECROW* (*SCR*) TF, so they together induce *miRNA165/166* expression (Carlsbecker et al., 2010; De Rybel et al., 2016). MicroRNAs diffuse into neighboring cells, creating a concentration gradient towards the center of the root. In the stele, *miRNA165/166* suppress the expression of the genes encoding the TFs of the *HD-ZIP III* family, limiting it to the central domain (see Fig. 7). In such a way, the metaxylem cells are predetermined. Whether this mechanism works in embryogenesis remains unknown, but this is a possibility since the *PHABULOSA* (*PHB*) TF of the *HD-ZIP III* family is expressed in the embryo root (Grigg et al., 2009).

Predetermination of phloem elements

The phloem markers expressed in progenitors and induce the tissue development include a number of the DOF family TFs (Miyashima et al., 2019; Roszak et al., 2021); strigolactone signaling pathway suppressors SUPPRESSOR OF MAX2 1-LIKE 3 (SMLX3), SMLX4 and SMLX5 (Wallner et al., 2017); membrane proteins BREVIS RADIX (BRX), OCTOPUS (OPS), OPS-LIKE 2 (OPL2) (Ruiz Sola et al., 2017); phosphatase COTYLEDON VASCULAR PATTERN 2 (CVP2) and its homolog CVP2-LIKE 1 (CVL1) (Rodriguez-Villalon et al., 2015); the ALTERED PHLOEM DEVELOPMENT (APL) TF (Bonke et al., 2003).

The formation of protophloem elements is controlled by shifting the balance towards inducing or suppressing mechanisms with the central link connecting the opposing regulatory modules being phloem-specific TFs of the DOF family (Qian et al., 2022). On the one hand, these TFs induce the expression of phloem development activators, such as *APL* as well as their own genes, forming a positive feedback loop. On the other hand, DOFs induce the expression of CLE25, CLE26, and CLE45 signaling peptides migrating to neighboring cells where they trigger an inhibitory regulatory module (see Fig. 7). Interacting with the BARELY ANY MERISTEM (BAM) receptors and the CLAVATA3 INSENSITIVE RECEPTOR KINASE (CIK) co-receptors, the CLE peptides induce the degradation of the DOF family TFs, suppressing the formation of protophloem elements. The activity of the CLE peptide receptors can be additionally regulated, e. g., by the MEMBRANE-ASSOCIATED KINASE REGULATOR 5 (MAKR5) (Kang, Hardtke, 2016) or CORYNE (CRN) (Hazak et al., 2017) regulators. The TFs of the DOF family activate the expression of the genes encoding the OPS membrane protein suppressing the BAM-CIK module (Qian et al., 2022). Properly positioned protophloem progenitor cells overcome the inhibitory effect of CLE peptides due to the DOF TF accumulation determined by the positive feedback. Such a balancing mechanism makes it possible to repattern the phloem in case protophloem development has been disrupted (Gujas et al., 2020). Here it should be noted that metaphloem development is probably regulated by other mechanisms and does not depend on that of the protophloem (Graeff, Hardtke, 2021).

During phloem formation, the phloem/procambium stem cell divides anticlinally to produce a daughter procambium and sieve-element progenitor to divide periclinally and form a procambium progenitor and a phloem sieve-element progenitor. The latter undergoes another periclinal division to produce proto- and metaphloem progenitors (Rodriguez-Villalon, 2016). Companion-like cells are another product of asymmetric division, but come from a different initial. These asymmetric cell divisions are controlled by a positional signal, a SHR-protein gradient whose migration into the endodermis activates miRNA165/166 and induces asymmetric divisions producing companion-like cells, while SHR movement into the phloem is necessary for the asymmetric divisions leading to proto- and metaxylem formation (Kim et al., 2020).

Conclusions

The vascular system of *A. thaliana* root is set at the earliest stages of embryogenesis. Wherein, the predetermination of provascular initials implies a labile, unstable, and reversible

specification based on the physical arrangement of cells in the embryo and influenced by a complex regulatory network of transcription factors. An interesting moment here is that both xylem (e. g., TMO5, T5L1) and phloem (e. g., PEAR1, TMO6, DOF6) markers are jointly expressed by provascular initials in early embryogenesis, but later they are separated into different spatial domains in the provascular meristem and seedling.

In *A. thaliana*, the vascular system is patterned by the time of embryo maturation. Partially, the gene network that controls this process in embryogenesis continues to maintain the vascular system structure of the growing root of the seedling and later during plant ontogenesis. This is associated with local accumulation of the molecular markers that are stably expressed in progenitor cells of a certain type. However, the factors working both in embryogenesis and during post-embryonic development can act at these stages in different ways.

Despite the significant progress that has recently been achieved in understanding the molecular and genetic mechanisms regulating vascular system development in plants, many questions remain open, in particular, those related to the existence of parallel regulatory pathways and feedforward loops. This is a good basis for building mathematical models whose analysis helps shed light on the relationship between various regulatory circuits and their functional significance.

References

- Agustí J., Blázquez M.A. Plant vascular development: mechanisms and environmental regulation. *Cell. Mol. Life Sci.* 2020;77(19):3711-3728. DOI 10.1007/s00018-020-03496-w.
- Bauby H., Divol F., Truernit E., Grandjean O., Palauqui J.C. Protophloem differentiation in early *Arabidopsis thaliana* development. *Plant Cell Physiol.* 2007;48(1):97-109. DOI 10.1093/pcp/pcl045.
- Baum S.F., Dubrovsky J.G., Rost T.L. Apical organization and maturation of the cortex and vascular cylinder in *Arabidopsis thaliana* (Brassicaceae) roots. *Am. J. Bot.* 2002;89(6):908-920. DOI 10.3732/ajb.89.6.908.
- Bishopp A., Help H., El-Showk S., Weijers D., Scheres B., Friml J., Benková E., Mähönen A.P., Helariutta Y. A mutually inhibitory interaction between auxin and cytokinin specifies vascular pattern in roots. *Curr. Biol.* 2011a;21(11):917-926. DOI 10.1016/j.cub.2011.04.017.
- Bishopp A., Lehesranta S., Vaten A., Help H., El-Showk S., Scheres B., Helariutta K., Mähönen A.P., Sakakibara H., Helariutta Y. Phloem-transported cytokinin regulates polar auxin transport and maintains vascular pattern in the root meristem. *Curr. Biol.* 2011b;21(11):927-932. DOI 10.1016/j.cub.2011.04.049.
- Bonke M., Thitamadee S., Mähönen A.P., Hauser M.T., Helariutta Y. APL regulates vascular tissue identity in *Arabidopsis*. *Nature.* 2003;426(6963):181-186. DOI 10.1038/nature02100.
- Busse J.S., Evert R.F. Pattern of differentiation of the first vascular elements in the embryo and seedling of *Arabidopsis thaliana*. *Int. J. Plant Sci.* 1999;160(1):1-13. DOI 10.1086/314098.
- Cai Q., Fukushima H., Yamamoto M., Ishii N., Sakamoto T., Kurata T., Motose H., Takahashi T. The SAC51 family plays a central role in thermopermine responses in *Arabidopsis*. *Plant Cell Physiol.* 2016;57(8):1583-1592. DOI 10.1093/pcp/pcw113.
- Carlsbecker A., Lee J.Y., Roberts C.J., Dettmer J., Lehesranta S., Zhou J., Lindgren O., Moreno-Risueno M.A., Vaten A., Thitamadee S., Campilho A., Sebastian J., Bowman J.L., Helariutta Y., Benfey P.N. Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature.* 2010;465(7296):316-321. DOI 10.1038/nature08977.

- Cheng C.Y., Kieber J.J. Cytokinin signaling in plants. In: Howell S. (Ed.) Molecular Biology. The Plant Sciences. Vol. 2. New York: Springer, 2014;269-289. DOI 10.1007/978-1-4614-7570-5_14.
- Courtois-Moreau C.L., Pesquet E., Sjödin A., Muñoz L., Bollhöner B., Kaneda M., Samuels L., Jansson S., Tuominen H. A unique program for cell death in xylem fibers of *Populus* stem. *Plant J.* 2009;58(2): 260-274. DOI 10.1111/j.1365-3113X.2008.03777.x.
- 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., Nijse B., Boekschooten 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*. 2014a;345(6197):1255215. DOI 10.1126/science.1255215.
- De Rybel B., Breda A.S., Weijers D. Prenatal plumbing – vascular tissue formation in the plant embryo. *Physiol. Plant.* 2014b;151(2): 126-133. DOI 10.1111/ppl.12091.
- De Rybel B., Mähönen A.P., Helariutta Y., Weijers D. Plant vascular development: from early specification to differentiation. *Nat. Rev. Mol. Cell Biol.* 2016;17(1):30-40. DOI 10.1038/nrm.2015.6.
- De Rybel B., Möller B., Yoshida S., Grabowicz I., Barbier de Reuille P., Boeren S., Smith R.S., Borst J.W., Weijers D. A bHLH complex controls embryonic vascular tissue establishment and indeterminate growth in *Arabidopsis*. *Dev. Cell.* 2013;24(4):426-437. DOI 10.1016/j.devcel.2012.12.013.
- Desvoves B., Echevarría C., Gutierrez C. A perspective on cell proliferation kinetics in the root apical meristem. *J. Exp. Bot.* 2021;72(19): 6708-6715. DOI 10.1093/jxb/erab303.
- Dolan L., Janmaat K., Willemsen V., Linstead P., Poethig S., Roberts K., Scheres B. Cellular organisation of the *Arabidopsis thaliana* root. *Development*. 1993;119(1):71-84. DOI 10.1242/dev.119.1.71.
- Evert R.F., Eichhorn S.E. Esau's Plant Anatomy. Meristems, Cells, and Tissues of the Plant Body: Their Structure, Function, and Development. New Jersey: Wiley, 2006. DOI 10.1002/0470047380.
- Friml J., Vieten A., Sauer M., Weijers D., Schwarz H., Hamann T., Offringa R., Jürgens G. Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature*. 2003;426(6963):147-153. DOI 10.1038/nature02085.
- Furuta K.M., Hellmann E., Helariutta Y. Molecular control of cell specification and cell differentiation during procambial development. *Annu. Rev. Plant. Biol.* 2014;65:607-638. DOI 10.1146/annurev-arplant-050213-040306.
- Graeff M., Hardtke C.S. Metaphloem development in the *Arabidopsis* root tip. *Development*. 2021;148(18):dev199766. DOI 10.1242/dev.199766.
- Grigg S.P., Galinha C., Kornet N., Canales C., Scheres B., Tsiantis M. Repression of apical homeobox genes is required for embryonic root development in *Arabidopsis*. *Curr. Biol.* 2009;19(17):1485-1490. DOI 10.1016/j.cub.2009.06.070.
- Gujas B., Kastanaki E., Sturchler A., Cruz T.M.D., Ruiz-Sola M.A., Dreos R., Eicke S., Truernit E., Rodriguez-Villalon A. A reservoir of pluripotent phloem cells safeguards the linear developmental trajectory of protophloem sieve elements. *Curr. Biol.* 2020;30(5):755-766. DOI 10.1016/j.cub.2019.12.043.
- Hazak O., Brandt B., Cattaneo P., Santiago J., Rodriguez-Villalon A., Hothorn M., Hardtke C.S. Perception of root-active CLE peptides requires CORYNE function in the phloem vasculature. *EMBO Rep.* 2017;18(8):1367-1381. DOI 10.15252/embr.201643535.
- Help H., Mähönen A.P., Helariutta Y., Bishopp A. Bisymmetry in the embryonic root is dependent on cotyledon number and position. *Plant Signal. Behav.* 2011;6(11):1837-1840. DOI 10.4161/psb.6.11.17600.
- Heo J.O., Roszak P., Furuta K.M., Helariutta Y. Phloem development: current knowledge and future perspectives. *Am. J. Bot.* 2014;101(9): 1393-1402. DOI 10.3732/ajb.1400197.
- Kang Y.H., Hardtke C.S. *Arabidopsis* MAK5 is a positive effector of BAM3-dependent CLE45 signaling. *EMBO Rep.* 2016;17(8):1145-1154. DOI 10.15252/embr.201642450.
- Katayama H., Iwamoto K., Kariya Y., Asakawa T., Kan T., Fukuda H., Ohashi-Ito K. A negative feedback loop controlling bHLH complexes is involved in vascular cell division and differentiation in the root apical meristem. *Curr. Biol.* 2015;25(23):3144-3150. DOI 10.1016/j.cub.2015.10.051.
- Kim H., Zhou J., Kumar D., Jang G., Ryu K.H., Sebastian J., Miyashima S., Helariutta Y., Lee J.Y. SHORTROOT-mediated intercellular signals coordinate phloem development in *Arabidopsis* roots. *Plant Cell.* 2020;32(5):1519-1535. DOI 10.1105/tpc.19.00455.
- Kuroha T., Tokunaga H., Kojima M., Ueda N., Ishida T., Nagawa S., Fukuda H., Sugimoto K., Sakakibara H. Functional analyses of *LONELY GUY* cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. *Plant Cell.* 2009; 21(10):3152-3169. DOI 10.1105/tpc.109.068676.
- Lau S., De Smet I., Kolb M., Meinhardt H., Jürgens G. Auxin triggers a genetic switch. *Nat. Cell Biol.* 2011;13(5):611-615. DOI 10.1038/ncb2212.
- Lau S., Slane D., Herud O., Kong J., Jürgens G. Early embryogenesis in flowering plants: setting up the basic body pattern. *Annu. Rev. Plant Biol.* 2012;63:483-506. DOI 10.1146/annurev-arplant-042811-105507.
- Li X., Wu H.X., Southerton S.G. Comparative genomics reveals conservative evolution of the xylem transcriptome in vascular plants. *BMC Evol. Biol.* 2010;10:190. DOI 10.1186/1471-2148-10-190.
- Lucas W.J., Groover A., Lichtenberger R., Furuta K., Yadav S.R., Helariutta Y., He X.Q., Fukuda H., Kang J., Brady S.M., Patrick J.W., Sperry J., Yoshida A., López-Millán A.F., Grusak M.A., Kachroo P. The plant vascular system: evolution, development and functions. *J. Integr. Plant Biol.* 2013;55(4):294-388. DOI 10.1111/jipb.12041.
- Mähönen A.P., Bishopp A., Higuchi M., Nieminen K.M., Kinoshita K., Törmäkangas K., Ikeda Y., Oka A., Kakimoto T., Helariutta Y. Cytokinin signaling and its inhibitor AHP6 regulate cell fate during vascular development. *Science*. 2006;311(5757):94-98. DOI 10.1126/science.1118875.
- Mähönen A.P., Bonke M., Kauppinen L., Riikonen M., Benfey P.N., Helariutta Y. A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.* 2000; 14(23):2938-2943. DOI 10.1101/gad.189200.
- Marhavý P., Bielach A., Abas L., Abuzeineh A., Duclercq J., Tanaka H., Páezová M., Petrášek J., Friml J., Kleine-Vehn J., Benková E. Cytokinin modulates endocytic trafficking of PIN1 auxin efflux carrier to control plant organogenesis. *Dev. Cell.* 2011;21(4):796-804. DOI 10.1016/j.devcel.2011.08.014.
- Ménard D., Pesquet E. Cellular interactions during tracheary elements formation and function. *Curr. Opin. Plant Biol.* 2015;23:109-115. DOI 10.1016/j.pbi.2014.12.001.
- Mironova V., Teale W., Shahriari M., Dawson J., Palme K. The systems biology of auxin in developing embryos. *Trends Plant Sci.* 2017; 22(3):225-235. DOI 10.1016/j.tplants.2016.11.010.
- Miyashima S., Roszak P., Seville I., Toyokura K., Blob B., Heo J.O., Mellor N., Help-Rinta-Rahko H., Otero S., Smet W., Boekschooten M., Hooiveld G., Hashimoto K., Smetana O., Siligato R., Wallner E.S., Mähönen A.P., Kondo Y., Melnyk C.W., Greb T., Nakajima K., Sozzani R., Bishopp A., De Rybel B., Helariutta Y. Mobile PEAR transcription factors integrate positional cues to prime cambial growth. *Nature*. 2019;565(7740):490-494. DOI 10.1038/s41586-018-0839-y.
- Miyashima S., Sebastian J., Lee J.Y., Helariutta Y. Stem cell function during plant vascular development. *EMBO J.* 2013;32(2):178-193. DOI 10.1038/emboj.2012.301.
- Möller B.K., ten Hove C.A., Xiang D., Williams N., López L.G., Yoshida S., Smit M., Datla R., Weijers D. Auxin response cell-autonomously controls ground tissue initiation in the early *Arabidopsis* embryo. *Proc. Natl. Acad. Sci. USA.* 2017;114(12):2533-2539. DOI 10.1073/pnas.1616493114.

- Möller B., Weijers D. Auxin control of embryo patterning. *Cold Spring Harb. Perspect. Biol.* 2009;1(5):a001545. DOI 10.1101/cshperspect.a001545.
- Muraro D., Mellor N., Pound M.P., Help H., Lucas M., Chopard J., Byrne H.M., Godin C., Hodgman T.C., King J.R., Pridmore T.P., Helariutta Y., Bennett M.J., Bishopp A. Integration of hormonal signaling networks and mobile microRNAs is required for vascular patterning in *Arabidopsis* roots. *Proc. Natl. Acad. Sci. USA.* 2014; 111(2):857-862. DOI 10.1073/pnas.1221766111.
- Nieminen K., Blomster T., Helariutta Y., Mähönen A.P. Vascular cambium development. *Arabidopsis Book.* 2015;13:e0177. DOI 10.1199/tab.0177.
- Ohashi-Ito K., Bergmann D.C. Regulation of the *Arabidopsis* root vascular initial population by *LONESOME HIGHWAY*. *Development.* 2007;134(16):2959-2968. DOI 10.1242/dev.006296.
- Ohashi-Ito K., Matsukawa M., Fukuda H. An atypical bHLH transcription factor regulates early xylem development downstream of auxin. *Plant Cell Physiol.* 2013;54(3):398-405. DOI 10.1093/pcp/pct013.
- Ohashi-Ito K., Saegusa M., Iwamoto K., Oda Y., Katayama H., Kojima M., Sakakibara H., Fukuda H. A bHLH complex activates vascular cell division via cytokinin action in root apical meristem. *Curr. Biol.* 2014;24(17):2053-2058. DOI 10.1016/j.cub.2014.07.050.
- Palovaara J., Saiga S., Wendrich J.R., van't Wout Hofland N., van Schayck J.P., Hater F., Mutte S., Sjollem J., Boekschoten M., Hooiveld G.J., Weijers D. Transcriptome dynamics revealed by a gene expression atlas of the early *Arabidopsis* embryo. *Nat. Plants.* 2017;3(11):894-904. DOI 10.1038/s41477-017-0035-3.
- Qian P., Song W., Zaizen-Iida M., Kume S., Wang G., Zhang Y., Kinoshita-Tsujimura K., Chai J., Kakimoto T. A Dof-CLE circuit controls phloem organization. *Nat. Plants.* 2022;8(7):817-827. DOI 10.1038/s41477-022-01176-0.
- Ren S.C., Song X.F., Chen W.Q., Lu R., Lucas W.J., Liu C.M. CLE25 peptide regulates phloem initiation in *Arabidopsis* through a CLERK-CLV2 receptor complex. *J. Integr. Plant. Biol.* 2019; 61(10):1043-1061. DOI 10.1111/jipb.12846.
- Robert H.S., Crhak Khaitova L., Mroue S., Benková E. The importance of localized auxin production for morphogenesis of reproductive organs and embryos in *Arabidopsis*. *J. Exp. Bot.* 2015;66(16):5029-5042. DOI 10.1093/jxb/erv256.
- Rodriguez-Villalon A. Wiring a plant: genetic networks for phloem formation in *Arabidopsis thaliana* roots. *New Phytol.* 2016;210(1): 45-50. DOI 10.1111/nph.13527.
- Rodriguez-Villalon A., Gujas B., van Wijk R., Munnik T., Hardtke C.S. Primary root protophloem differentiation requires balanced phosphatidylinositol-4,5-bisphosphate levels and systemically affects root branching. *Development.* 2015;142(8):1437-1446. DOI 10.1242/dev.118364.
- Ross-Elliott T.J., Jensen K.H., Haaning K.S., Wager B.M., Knoblauch J., Howell A.H., Mullendore D.L., Monteith A.G., Paultre D., Yan D., Otero S., Bourdon M., Sager R., Lee J.Y., Helariutta Y., Knoblauch M., Oparka K.J. Phloem unloading in *Arabidopsis* roots is convective and regulated by the phloem-pole pericycle. *eLife.* 2017;6:e24125. DOI 10.7554/eLife.24125.
- Roszak P., Heo J.O., Blob B., Toyokura K., Sugiyama Y., de Luis Balaguer M.A., Lau W.W.Y., Hamey F., Cirrone J., Madej E., Bouatta A.M., Wang X., Guichard M., Ursache R., Tavares H., Verstaen K., Wendrich J., Melnyk C.W., Oda Y., Shasha D., Ahnert S.E., Saey Y., De Rybel B., Heidstra B., Scheres B., Grossmann G., Mähönen A.P., Denninger P., Göttgens B., Sozzani R., Birnbaum K.D., Helariutta Y. Cell-by-cell dissection of phloem development links a maturation gradient to cell specialization. *Science.* 2021;374(6575):eaba5531. DOI 10.1126/science.aba5531.
- Ruiz Sola M.A., Coiro M., Crivelli S., Zeeman S.C., Hansen S.S.K., Truernit E. *OCTOPUS-LIKE 2*, a novel player in *Arabidopsis* root and vascular development, reveals a key role for *OCTOPUS* family genes in root metaploem sieve tube differentiation. *New Phytol.* 2017;216(4):1191-1204. DOI 10.1111/nph.14751.
- Růžicka K., Ursache R., Hejácí J., Helariutta Y. Xylem development – from the cradle to the grave. *New Phytol.* 2015;207(3):519-535. DOI 10.1111/nph.13383.
- Sauer M., Balla J., Luschnig C., Wisniewska J., Reinöhl V., Friml J., Benková E. Canalization of auxin flow by Aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes Dev.* 2006;20(20):2902-2911. DOI 10.1101/gad.390806.
- Scarpella E., Meijer A.H. Pattern formation in the vascular system of monocot and dicot plant species. *New Phytol.* 2004;164(2):209-242. DOI 10.1111/j.1469-8137.2004.01191.x.
- Scheres B., Wolkenfelt H., Willemsen V., Terlouw M., Lawson E., Dean C., Weisbeek P. Embryonic origin of the *Arabidopsis* primary root and root meristem initials. *Development.* 1994;120(9):2475-2487. DOI 10.1242/dev.120.9.2475.
- Schlereth A., Möller B., Liu W., Kientz M., Flipse J., Rademacher E.H., Schmid M., Jürgens G., Weijers D. MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. *Nature.* 2010;464(7290):913-916. DOI 10.1038/nature08836.
- Seo M., Kim H., Lee J.Y. Information on the move: vascular tissue development in space and time during postembryonic root growth. *Curr. Opin. Plant Biol.* 2020;57:110-117. DOI 10.1016/j.pbi.2020.08.002.
- Sjolund R.D. The phloem sieve element: a river runs through it. *Plant Cell.* 1997;9(7):1137-1146. DOI 10.1105/tpc.9.7.1137.
- Smet W., Seville I., de Luis Balaguer M.A., Wybouw B., Mor E., Miyashima S., Blob B., Roszak P., Jacobs T.B., Boekschoten M., Hooiveld G., Sozzani R., Helariutta Y., De Rybel B. DOF2.1 controls cytokinin-dependent vascular cell proliferation downstream of TMO5/LHW. *Curr. Biol.* 2019;29(3):520-529.e6. DOI 10.1016/j.cub.2018.12.041.
- Smetana O., Mäkilä R., Lyu M., Amiryousefi A., Sánchez Rodríguez F., Wu M.-F., Solé-Gil A., Leal Gavarrón M., Siligato R., Miyashima S., Roszak P., Blomster T., Reed J.W., Broholm S., Mähönen A.P. High levels of auxin signalling define the stem-cell organizer of the vascular cambium. *Nature.* 2019;565(7740):485-489. DOI 10.1038/s41586-018-0837-0.
- Smit M.E., Llavata-Peris C.I., Roosjen M., van Beijnum H., Novikova D., Levitsky V., Seville I., Roszak P., Slane D., Jürgens G., Mironova V., Brady S.M., Weijers D. Specification and regulation of vascular tissue identity in the *Arabidopsis* embryo. *Development.* 2020;147(8):dev186130. DOI 10.1242/dev.186130.
- Smit M.E., Weijers D. The role of auxin signaling in early embryo pattern formation. *Curr. Opin. Plant Biol.* 2015;28:99-105. DOI 10.1016/j.pbi.2015.10.001.
- Smith R.A., Schuetz M., Roach M., Mansfield S.D., Ellis B., Samuels L. Neighboring parenchyma cells contribute to *Arabidopsis* xylem lignification, while lignification of interfascicular fibers is cell autonomous. *Plant Cell.* 2013;25(10):3988-3999. DOI 10.1105/tpc.113.117176.
- Stadler R., Wright K.M., Lauterbach C., Amon G., Gahrz M., Feuerstein A., Oparka K.J., Sauer N. Expression of GFP-fusions in *Arabidopsis* companion cells reveals non-specific protein trafficking into sieve elements and identifies a novel post-phloem domain in roots. *Plant J.* 2005;41(2):319-331. DOI 10.1111/j.1365-313X.2004.02298.x.
- Tanaka H., Dhonukshe P., Brewer P.B., Friml J. Spatiotemporal asymmetric auxin distribution: a means to coordinate plant development. *Cell. Mol. Life Sci.* 2006;63(23):2738-2754. DOI 10.1007/s00018-006-6116-5.
- Truernit E. Sieve elements and their cell neighbours in the *Arabidopsis* root – roles and relationships. *J. Plant Physiol.* 2022;268:153569. DOI 10.1016/j.jplph.2021.153569.
- Tvorogova V.E., Osipova M.A., Doduyeva I.E., Lutova L.A. Interaction between transcriptional factors and phytohormones in regulation of plant meristems activity. *Ekologicheskaya Genetika = Ecological Genetics.* 2012;10(3):28-40. (in Russian)
- Vera-Sirera F., De Rybel B., Úrbez C., Kouklas E., Pesquera M., Álvarez-Mahecha J.C., Minguet E.G., Tuominen H., Carbonell J.,

- Borst J.W., Weijers D., Blázquez M.A. A bHLH-based feedback loop restricts vascular cell proliferation in plants. *Dev. Cell.* 2015; 35(4):432-443. DOI 10.1016/j.devcel.2015.10.022.
- Verma S., Attuluri V.P.S., Robert H.S. An essential function for auxin in embryo development. *Cold Spring Harb. Perspect. Biol.* 2021; 13(4):a039966. DOI 10.1101/cshperspect.a039966.
- Wabnik K., Robert H.S., Smith R.S., Friml J. Modeling framework for the establishment of the apical-basal embryonic axis in plants. *Curr. Biol.* 2013;23(24):2513-2518. DOI 10.1016/j.cub.2013.10.038.
- Wallner E.S., López-Salmerón V., Belevich I., Poschet G., Jung I., Grünwald K., Sevilem I., Jokitalo E., Hell R., Helariutta Y., Agustí J., Lebovka I., Greb T. Strigolactone- and karrikin-independent SMXL proteins are central regulators of phloem formation. *Curr. Biol.* 2017;27(8):1241-1247. DOI 10.1016/j.cub.2017.03.014.
- Weijers D., Jürgens G. Auxin and embryo axis formation: the ends in sight? *Curr. Opin. Plant Biol.* 2005;8(1):32-37. DOI 10.1016/j.pbi.2004.11.001.
- Yang B., Minne M., Brunoni F., Plačková L., Petřík I., Sun Y., Nolf J., Smet W., Verstaen K., Wendrich J.R., Eekhout T., Hoyerová K., van Isterdael G., Hastraete J., Bishopp A., Farcot E., Novák O., Saeys Y., de Rybel B. Non-cell autonomous and spatiotemporal signalling from a tissue organizer orchestrates root vascular development. *Nat. Plants.* 2021;7(11):1485-1494. DOI 10.1038/s41477-021-01017-6.

Acknowledgements. The study was supported by Russian Government Project FWN-2022-0020. The microscopy analysis was done in the frame of the project supported by the Russian Science Foundation, grant No. 20-14-00140. The authors thank Joint Access Center for Microscopy of Biological Objects, SB RAS (<http://www.bionet.nsc.ru/microscopy/>) for granting access to microscopic equipment. The authors are deeply grateful to Prof. Dolf Weijers for kindly providing *Arabidopsis thaliana* reporter lines.

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

Received September 16, 2022. Revised November 8, 2022. Accepted November 10, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins

E.A. Antropova¹✉, T.M. Khlebodarova^{1, 2}, P.S. Demenkov^{1, 2}, A.S. Venzel^{1, 2}, N.V. Ivanisenko^{1, 2}, A.D. Gavrilenko^{1, 3}, T.V. Ivanisenko^{1, 2}, A.V. Adamovskaya^{2, 3}, P.M. Revva^{2, 3}, I.N. Lavrik⁴, 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 Translational Inflammation Research, Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany

✉ nzhenia@bionet.nsc.ru

Abstract. Hepatitis C virus (HCV) is a risk factor that leads to hepatocellular carcinoma (HCC) development. Epigenetic changes are known to play an important role in the molecular genetic mechanisms of virus-induced oncogenesis. Aberrant DNA methylation is a mediator of epigenetic changes that are closely associated with the HCC pathogenesis and considered a biomarker for its early diagnosis. The ANDSystem software package was used to reconstruct and evaluate the statistical significance of the pathways HCV could potentially use to regulate 32 hypermethylated genes in HCC, including both oncosuppressor and protumorigenic ones identified by genome-wide analysis of DNA methylation. The reconstructed pathways included those affecting protein-protein interactions (PPI), gene expression, protein activity, stability, and transport regulations, the expression regulation pathways being statistically significant. It has been shown that 8 out of 10 HCV proteins were involved in these pathways, the HCV NS3 protein being implicated in the largest number of regulatory pathways. NS3 was associated with the regulation of 5 tumor-suppressor genes, which may be the evidence of its central role in HCC pathogenesis. Analysis of the reconstructed pathways has demonstrated that following the transcription factor inhibition caused by binding to viral proteins, the expression of a number of oncosuppressors (*WT1*, *MGMT*, *SOCS1*, *P53*) was suppressed, while the expression of others (*RAS1*, *RUNX3*, *WIF1*, *DAPK1*) was activated. Thus, the performed gene-network reconstruction has shown that HCV proteins can influence not only the methylation status of oncosuppressor genes, but also their transcriptional regulation. The results obtained can be used in the search for pharmacological targets to develop new drugs against HCV-induced HCC.

Key words: hepatocellular carcinoma; hepatitis C virus; expression regulation; methylation; regulatory pathways; gene networks; bioinformatics.

For citation: 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 Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. DOI 10.18699/VJGB-22-89

Компьютерный анализ особенностей регуляции гиперметилированных маркерных генов гепатокарциномы вирусными белками гепатита С

Е.А. Антропова¹✉, Т.М. Хлебодарова^{1, 2}, П.С. Деменков^{1, 2}, А.С. Вензель^{1, 2}, Н.В. Иванисенко^{1, 2}, А.Д. Гавриленко^{1, 3}, Т.В. Иванисенко^{1, 2}, А.В. Адамовская^{2, 3}, П.М. Ревва^{2, 3}, И.Н. Лаврик⁴, 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 Магдебургский университет им. Отто фон Герике, медицинский факультет, Магдебург, Германия

✉ nzhenia@bionet.nsc.ru

Аннотация. Вирус гепатита С (ВГС) считается фактором риска для возникновения гепатоцеллюлярной карциномы (ГЦК). Известно, что большую роль в молекулярно-генетических механизмах вирус-индуцированного онкогенеза играют эпигенетические изменения. Аберрантное метилирование ДНК служит медиатором эпигенетических изменений, которые тесно связаны с патогенезом ГЦК, и признано биомаркером для его ранней диагностики. С помощью ANDSystem проведены реконструкция и оценка статистической значимости путей потенциальной регуляции вирусными белками ВГС 32 генов человека, гиперметилированных при ГЦК. Среди исследованных генов были как онкосупрессоры, так и проопухольевые гены, идентифицированных по данным

полногеномного анализа метилирования ДНК. Реконструированы регуляторные пути, включающие белок-белковые взаимодействия, регуляцию экспрессии генов, регуляцию активности, стабильности и транспорта белков. Среди статистически значимых оказались пути регуляции экспрессии. Показано, что восемь из десяти белков ВГС являются участниками данных путей. Белок ВГС NS3 был вовлечен в наибольшее число регуляторных путей. NS3 связан с регуляцией пяти генов-онкосупрессоров, что может свидетельствовать о его центральной роли в патогенезе ГЦК. Анализ реконструированных путей показал, что при ингибировании транскрипционных факторов в результате связывания с вирусными белками, экспрессия ряда онкосупрессоров (*WT1*, *MGMT*, *SOCS1*, *P53*) подавлялась, тогда как экспрессия других (*RAS1*, *RUNX3*, *WIF1*, *DAPK1*) активировалась. Таким образом, с помощью реконструкции генных сетей показано, что вирусные белки гепатита С способны влиять не только на статус метилирования генов-онкосупрессоров, но и на их транскрипционную регуляцию. Полученные результаты могут быть использованы при поиске фармакологических мишеней для разработки новых средств против ГЦК, индуцированной ВГС.

Ключевые слова: гепатоцеллюлярная карцинома; вирус гепатита С; регуляция экспрессии; гиперметилирование; регуляторные пути; генные сети; биоинформатика.

Introduction

Liver cancer is the third leading cause of cancer-related death in the world according to year 2020 statistics with over 900,000 new cases of this pathology registered the same year around the world (International Agency for Research on Cancer, <https://gco.iarc.fr/today/home>). Hepatocellular carcinoma (HCC) has been the dominant type of primary liver cancer, comprising about 90 % of all the cases (Llovet et al., 2016). It may be caused by several risk factors such as aflatoxin exposure, alcohol consumption; hepatitis B or C (HCV) virus infection, liver cirrhosis, non-alcoholic fatty liver disease, non-alcoholic steatohepatitis, metabolic syndrome, obesity, type II diabetes, and genetic predisposition (McGlynn et al., 2021).

Currently, a lot of data has been accumulated on HCV association with impaired liver function, cirrhosis and HCC development (Rabaan et al., 2020). Having gotten into a human body, HCV seeks to exercise control over the biological processes occurring in host cells in order to increase its survival and replication efficiency. In more than 70 % of those initially infected, the disease takes on a chronic course, so the patients experience progressive liver-tissue fibrosis and cirrhosis accompanied by long-term inflammation (Jaroszewicz et al., 2015). Using various mechanisms for infected cell cooptation, the virus can inadvertently lead to HCC development (D'souza et al., 2020). At the same time, the molecular and genetic mechanisms of virus-induced carcinogenesis remain understudied.

In addition, HCC pathogenesis is associated with epigenetic modifications and aberrant DNA methylation being a mediator of epigenetic changes (Fernández-Barrena et al., 2020) that can serve as a biomarker for early HCC diagnosis (Zhang C. et al., 2016; Xu et al., 2017).

To establish the functional links between genes and to elucidate the molecular mechanisms of biological processes, the methods for gene networks reconstruction have been widely employed. Previously, we developed the Associative Network Discovery System (ANDSystem) software package designed to reconstruct gene networks based on the knowledge extracted from factual databases and scientific publications using text-mining techniques (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020). The package has enabled one to reconstruct the molecular mechanisms of a number of pathologies such as preeclampsia (Glotov et al., 2015), tuberculosis

(Bragina et al., 2016), comorbid conditions of asthma and hypertension (Saik et al., 2018), COVID-19 (Ivanisenko N.V. et al., 2020), HCV life cycle (Saik et al., 2016), etc.

In the present study, ANDSystem was employed to reconstruct the regulatory pathways describing the potential regulation mechanisms of the genes hypermethylated in HCC by HCV proteins. The analysis looked at the 32 genes known to be hypermethylated HCC markers. Among the 7 types of reconstructed regulatory pathways including protein-protein interactions (PPI), gene expression, protein activity, stability and transport regulations, those responsible for gene expression regulation turned out to be statistically significant. Nine marker genes were identified that could potentially be subject to regulation by HCV proteins, including three HCC suppressor genes (*MGMT*, *SOCS1* and *TP53*) that could be negatively regulated and one apoptosis suppressor gene (*TERT*) that can be positively regulated.

Materials and methods

Genes hypermethylated in HCC. Information about the hypermethylated genes was taken from publications (Table 1). Only those genes were considered whose hypermethylation was associated with HCC and confirmed through the analysis and meta-analysis given in the publications. The schematic of the data-processing algorithm can be seen in Figure 1.

Regulatory pathways reconstruction in ANDSystem.

The regulatory pathways were reconstructed using the ANDSystem software package (Ivanisenko V.A. et al., 2019) that had been designed to perform gene-networks reconstruction based on automated analysis of scientific texts and factual databases. ANDSystem includes a knowledge base with more than 40 million facts about molecular-genetic interactions, containing physical intermolecular interactions, gene expression, protein activity, stability and transport regulations. In the package, it is the the ANDVisio program that reconstructs and analyzes gene networks using the Pathway Wizard function performing search calls to the knowledge base according to a given pattern. A schematic description of the used patterns is given in Table 2.

For instance, P4 means searching for all possible molecular genetic pathways in the ANDSystem knowledge base that satisfy the following requirement: the first participant in the pathway is the viral protein (Vp); the second is human protein (Hp); the third is a human gene from a list of target

Table 1. List of the hypermethylated genes used in the analysis

Gene	Encoded protein	Source
<i>APC</i>	Adenomatous polyposis coli	Zhang C. et al., 2016
<i>COL7A1</i>	Collagen alpha-1(VII) chain	Schulze et al., 2015
<i>COL16A1</i>	Collagen alpha-1(XVI) chain	Schulze et al., 2015
<i>DAPK1</i>	Death-associated protein kinase 1	Zhang C. et al., 2016
<i>DSE</i>	Dermatan-sulfate epimerase	Cheng et al., 2018
<i>FAM55C</i>	NXPE family member 3	Cheng et al., 2018
<i>FAT4</i>	Protocadherin Fat 4	Schulze et al., 2015
<i>GALNT3</i>	Polypeptide N-acetylgalactosaminyltransferase 3	Cheng et al., 2018
<i>GSTP1</i>	Glutathione S-transferase P	Zhang C. et al., 2016
<i>IGFALS</i>	Insulin-like growth-factor-binding protein, acid labile subunit	Neumann et al., 2012
<i>KCNA3</i>	Potassium voltage-gated channel subfamily A member 3	Hernandez-Meza et al., 2021
<i>LDHB</i>	L-lactate dehydrogenase B chain	Hernandez-Meza et al., 2021
<i>MGMT</i>	O6-methylguanine-DNA methyltransferase	Zhang C. et al., 2016
<i>NEBL</i>	Nebulette	Cheng et al., 2018
<i>NEFH</i>	Neurofilament heavy polypeptide	Revill et al., 2013
<i>OPCML</i>	Opioid binding protein/cell adhesion molecule-like	Zhang C. et al., 2016
<i>TP53</i>	Cellular tumor antigen p53	Zhang C. et al., 2016
<i>PER3</i>	Period circadian protein homolog 3	Neumann et al., 2012
<i>PRDM2</i>	PR domain containing 2	Zhang C. et al., 2016
<i>PROZ</i>	Vitamin K-dependent protein Z	Neumann et al., 2012
<i>RARβ</i>	Retinoic acid receptor beta	Zhang C. et al., 2016
<i>RASSF1A</i>	Ras association domain-containing protein 1	Zhang C. et al., 2016
<i>RUNX3</i>	Runt-related transcription factor 3	Zhang C. et al., 2016
<i>SFRP1</i>	Secreted frizzled-related protein 1	Zhang C. et al., 2016
<i>SMPD3</i>	Sphingomyelin phosphodiesterase 3	Revill et al., 2013
<i>SOCS1</i>	Suppressor of the cytokine signalling 1	Zhang C. et al., 2016
<i>SPINT2</i>	Serine peptidase inhibitor, Kunitz type, 2	Zhang C. et al., 2016; Hernandez-Meza et al., 2021
<i>TERT</i>	Telomerase reverse transcriptase	Zhang H. et al., 2015
<i>TSPYL5</i>	Testis-specific Y-encoded-like protein 5	Hernandez-Meza et al., 2021
<i>TTC36</i>	Tetratricopeptide repeat protein 36	Jing et al., 2022
<i>WIF1</i>	WNT inhibitory factor 1	Zhang C. et al., 2016
<i>WT1</i>	Wilms tumor 1	Zhang C. et al., 2016

genes (Tg); the last member of the pathway is a Tg-encoded protein (Tp). Further in the text, HCC marker genes will be regarded as target ones. Interactions between pathway participants are represented by the following types: Vp and Hp are linked by *PPIs*; Hp and Tg are “expression regulation” interaction type (*Exp reg*), where Hp regulates Tg gene expression; Tg and Tp are interaction of the “expression” interaction type (*Exp*), i.e., the Tp protein is the expression product of the Tg gene. Examples of regulatory pathway reconstruction

in ANDSystem using the patterns presented in the previous work (Ivanisenko V.A. et al., 2022).

Estimating the statistical significance of pathways. The patterns from Table 2 were used to calculate the number of marker genes *K* participating in the regulatory pathways, as well as the number of such participants in the sample of control gene. The likelihood of observing the number *K* for random reasons was estimated using the standard hypergeometric distribution and the hypergeom function from the

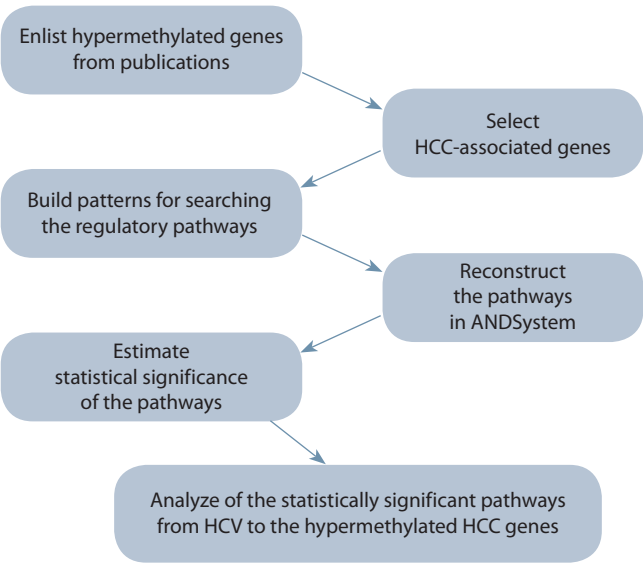


Fig. 1. Schematic description of the data processing algorithm.

SciPy 1.8.0 package (<https://scipy.org>). For the purposes of statistical processing, a group of genes proposed by Hoshida et al. (2008) as a control to predict HCC outcomes based on the expression level of genes was taken.

Results

Reconstruction of the potential regulatory pathways
HCV proteins use to affect HCC marker genes

A set of hypermethylated HCC marker genes was used to reconstruct the potential regulatory pathways through which viral proteins could modulate the genes playing an important role in HCC pathogenesis (see Table 1). The set had been based on the published results of a genome-wide analysis of DNA methylation and included 30 genes, the expression of which, according to the studies, was reduced in hepatocellular carcinoma, and two genes (*WT1* and *TERT*) with increased expression.

To reconstruct the regulatory pathways, the ANDSystem software package was used. The search queries to the knowledge base were formed using the pathway patterns presented in Table 2. The patterns described different types of regulatory pathways determined by different combinations of molecular-genetic interactions, including PPIs, gene expression, protein activity, stability, and transport regulations.

Analysis of the statistical significance of the pathways automatically reconstructed by ANDSystem according to the given patterns showed that among the seven types of regulatory pathways analyzed, expression regulation ones turned out to be statistically significant (P4 in Table 3). This pattern describes the pathways including four participants: (1) viral proteins; (2) human transcription factors (TF) involved in PPIs with viral proteins; (3) marker genes presented in Table 1, whose expression is regulated by (2); (4) protein products of marker genes.

The gene network describing the regulation pathways of HCC marker genes included 8 HCV proteins, 7 intermediate host proteins involved in PPIs with HCV proteins, and 9 genes

Table 2. Patterns to search for the regulatory pathways describing viral – protein modulation of HCC marker genes

Pattern #	Pattern scheme
P1	Vp \xrightarrow{PPI} Tp
P2	Vp \xrightarrow{PPI} Hp \xrightarrow{PPI} Tp
P3	Vp \xrightarrow{PPI} Hp $\xrightarrow{Act/Stab/Pr/PPM/Tr}$ Tp
P4	Vp \xrightarrow{PPI} Hp $\xrightarrow{Exp\ reg}$ Tg \xrightarrow{Exp} Tp
P5	Vp \xrightarrow{PPI} Hp $\xrightarrow{Exp\ reg}$ Hg \xrightarrow{Exp} Hp $\xrightarrow{Exp\ reg}$ Tg \xrightarrow{Exp} Tp
P6	Vp \xrightarrow{PPI} Hp $\xrightarrow{Exp\ reg}$ Hg \xrightarrow{Exp} Hp $\xrightarrow{Act/Stab/Pr/PPM/Tr}$ Tp
P7	Vp \xrightarrow{PPI} Hp $\xrightarrow{Exp\ reg}$ Hg \xrightarrow{Exp} Hp \xrightarrow{PPI} Tp

Note. Vp – HCV proteins; Hp – any human proteins involved in the interactions; Hg – any human genes involved in the interactions; Tg – target genes (HCC marker genes); Tp – Tg-encoded target genes; PPI – protein-protein interactions; Act/Stab/Pr/PPM/Tr – activity or stability regulation, or proteolysis, or post-translational modifications, or transport regulation function; Exp reg – gene-expression regulation; Exp – protein-producing gene expression.

Table 3. Results of assessing the significance of the regulatory pathways described by different patterns

Regulatory pathway pattern	Number of participating gene-markers	P-val	FDR
P1	0	–	–
P2	15	0.52	0.62
P3	5	0.17	0.34
P4	9	0.0054	0.032
P5	21	0.047	0.14
P6	10	0.39	0.58
P7	23	0.83	0.83

Note. P-val is the level of statistical significance; FDR is the level of statistical significance accounting for multiple comparisons as per the false discovery rate (expected proportion of false rejections).

(*DAPK1*, *SOCS1*, *MGMT*, *RASSF1*, *RUNX3*, *TP53*, *WIF1*, *WT1*, and *TERT*) whose aberrant expression correlated with HCC progression (Fig. 2).

Analysis of the regulatory pathways

The regulatory pathways involved 8 out of 10 HCV proteins and 6 human genes, which protein products acting as the intermediate participants the viral proteins could form protein heterocomplexes with. The latter included such genes of transcription factors as STAT3 (Signal transducer and activator of transcription 3), NR4A1 (Nuclear receptor subfamily 4 group A member 1), JUN (c-Jun/activator protein 1), BCL6 (B-cell lymphoma 6 protein), transmembrane receptor NOTC1 (Neurogenic locus notch homolog protein 1) and histone methyltransferase SMYD3 (Lysine methyltransferase SET and MYND domain containing protein 3).

Most of the viral proteins were associated with *RUNX3* and *WT1* regulation. Six of them (NS4A, Core, p23, gp32, NS1, and NS5B) interacted with NR4A1 being a general expression regulator of these two HCC marker genes.

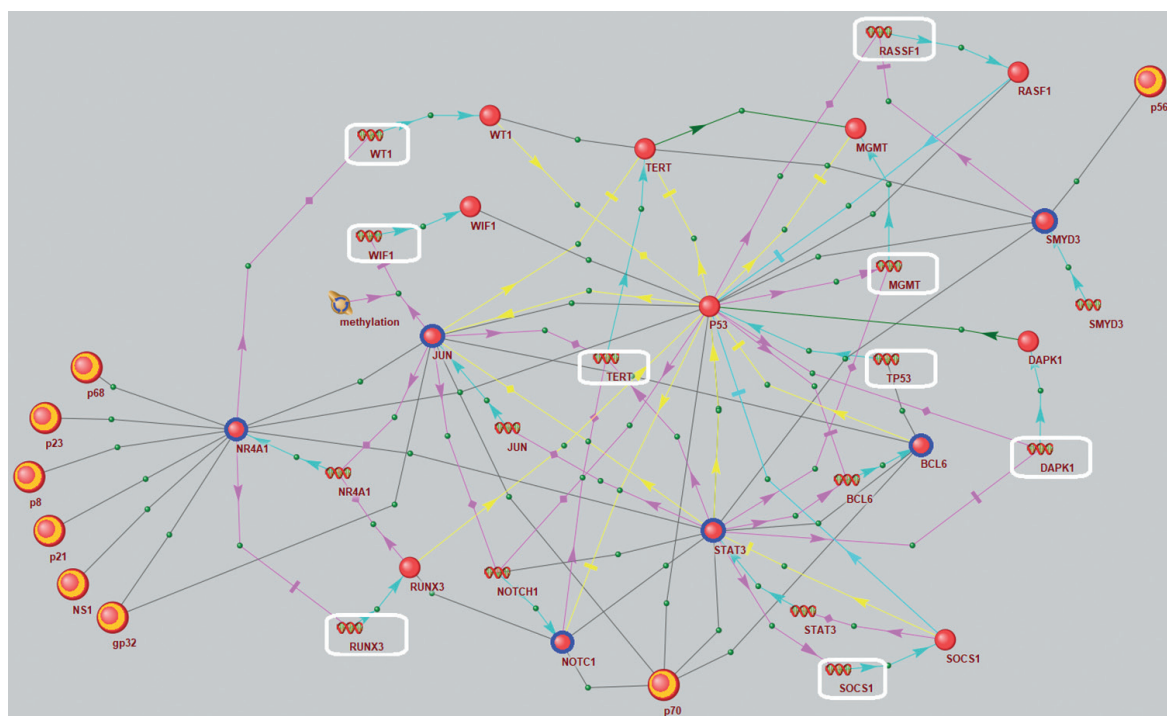


Fig. 2. Gene network including statistically significant regulatory pathways for the viral proteins to influence HCC marker genes expression that was reconstructed in ANDSystem using the P4 pattern.

Legend: HCV proteins (yellow-red large balls) – p8 (Non-structural protein 4A, NS4A), p21 (Core, Capsid protein C), p23 (Protease NS2-3), gp32 (Envelope glycoprotein E1), NS1 (Envelope glycoprotein E2), p56 (NS5A), p68 (NS5B), p70 (Hepacivirin, NS3); intermediate proteins (blue-red balls) – BCL6 (B-cell lymphoma 6 protein), NOTC1 (Neurogenic locus notch homolog protein 1), NR4A1 (Nuclear receptor subfamily 4 group A member 1), JUN (c-Jun/activator protein 1), SMYD3 (lysine methyltransferase SET and MYND domain containing protein 3), STAT3 (Signal transducer and activator of transcription 3); hypermethylated genes (highlighted in white frames) and their protein products – DAPK1 (Death associated protein kinase 1), MGMT (Methylated-DNA-protein-cysteine methyltransferase), RASSF1 (Ras association domain family member 1), RUNX3 (Runt-related transcription factor 3), SOCS1 (Suppressor of cytokine signaling 1), TERT (Telomerase reverse transcriptase), TP53 (Tumor protein p53), WIF1 (Wnt inhibitory factor 1), WT1 (Wilms tumor protein).

HCV protein NS3 (p70) interacted with the largest number of expression regulators, and through these interactions it could potentially regulate the expression of five tumor suppressor genes and that of *TERT*.

Now, let us consider the possibilities of implementing of these regulatory pathways in more detail.

p8, p21, p68, gp32, p23, NS1/NR4A1/RUNX3, WT1. This regulatory pathway suggests six HCV proteins (p8, p21, p68, gp32, p23, NS1) can possibly affect HCC development by controlling the activity of the *RUNX3* and *WT1* genes through the NR4A1 transcription factor. Indeed, NR4A1 directly interacts with the *RUNX3* and *WT1* promoters, suppressing *RUNX3* activity and activating that of *WT1* (Nowyhed et al., 2015; Zong et al., 2017). Both factors are involved in apoptosis regulation, hence, RUNX3 promotes activation of the extrinsic, TRAIL-induced apoptosis pathway (Kim et al., 2019), while WT1 controls the mitochondrial (internal) apoptosis pathway through the regulation of the *Bcl-2* anti-apoptotic protein gene, and, depending on a cell type, affects the expression of the *Bcl-2* gene in both positive and negative ways (Mayo et al., 1999; Loeb, 2006). It has been shown that in HCC, an increased expression of the *WT1* gene is observed, which is due to hypermethylation of its promoter and correlates with a poor prognosis (Sera et al., 2008; Mžik et al., 2016). These data suggest that the role WT1 plays in HCC progression is associated with blocked apoptosis.

Experiments have demonstrated that HCV core protein inhibits at least the *NR4A1* and *RUNX3* genes expression in infected cells (Tan, Li, 2015), contributing to suppressing an external apoptosis pathway. The Y2H test (Two Hybrid Test) has shown NR4A1 can interact with such viral proteins as CORE, E1, E2, NS2, NS4A, and NS5B (de Chassey et al., 2008), but except for CORE, the effects of the other HCV proteins on TF activity have not been investigated yet.

E1, NS3, Core, p23, NS1, p68/JUN, NOTC1, STAT3/TERT. Aberrant expression of the *TERT* gene, associated, among other things, with hypermethylation of its promoter, is a prognostic marker of HCC (Zhang H. et al., 2015; Zucman-Rossi et al., 2015; Oversoe et al., 2020). TERT affects disease progression through stimulation of cell proliferation due to reactivation of its gene expression in carcinoma cells (Nault et al., 2019; In der Stroth et al., 2020). TERT activity has been shown to also increase in HCV-infected cells, partly through direct interaction of the core protein with the enzyme (Zhu et al., 2010, 2017), however, in general, the mechanisms HCV proteins affect TERT activity are not clear. This regulatory pathway suggests the involvement of the virus NS3, Core, E1, p23, NS1, and NS5B proteins in *TERT* gene expression through interaction with the JUN (AP-1), STAT3, and NOTC1 proteins.

Indeed, experiments have shown that there is a possibility to affect *TERT* expression through the AP-1 and STAT3 TFs

being its direct regulators (Konnikova et al., 2005; Takakura et al., 2005), as well as through the NOTC1 signaling pathway (Sawangarun et al., 2018). Moreover, it has been demonstrated that the HCV NS3 protein affects NOTC1 activity through the SRCAP transcription factor (Iwai et al., 2011) and the expression of AP1- and STAT3-regulated genes (Hassan et al., 2005, 2007; Machida et al., 2006; Li et al., 2010), however, the specific mechanisms of realization of these influences in infected hepatocyte cells have not been practically examined.

gp32, p70/JUN/WIF1. This regulatory pathway describes the effect the NS3 and E1 HCV proteins have on *WIF1* (Wnt inhibitory factor 1) gene expression through interaction with the c-Jun/AP-1 TF. *WIF1* is a tumor suppressor that reduces cell growth in HCC (Deng et al., 2010), and its expression level is a prognostic indicator of the course of the disease (Huang et al., 2011).

Experiments have demonstrated that there is both the possibility of a direct effect of the NS3 and E1 proteins on the activity of c-Jun/AP-1 (de Chassesey et al., 2008), and the latter can affect the expression of the *WIF1* gene through interaction with the DNMT1 methyltransferase (DNA methyltransferase 1), suppressing *WIF1* in gallbladder cancer cells (Lin et al., 2018). But what mechanisms of *WIF1* gene suppression are initiated in HCV-infected hepatocarcinoma cells remains unknown.

p70/STAT3/MGMT, DAPK1, SOCS1. This regulatory pathway is initiated by NS3 (p70) affecting the activity of the *MGMT*, *SOCS*, and *DAPK1* genes through interaction with the STAT3 TF. Proteins DAPK1 (Death-associated protein kinase 1), *MGMT* (Methylated-DNA-protein-cysteine methyltransferase), and *SOCS1* (Suppressor of cytokine signaling 1) are considered tumor suppressors, and their low expression in carcinomas correlates with disease progression (Gui et al., 2011; Jiang et al., 2019; Chen J. et al., 2020; Chen P. et al., 2020; Song et al., 2020).

Experiments have demonstrated that NS3 can directly interact with the STAT3 TF (de Chassesey et al., 2008) involved in the regulation of the expression of the abovementioned genes (Kohsaka et al., 2012; Benderska, Schneider-Stock, 2014; Yang C. et al., 2015), however, the effect of STAT3 on *MGMT*, *SOCS*, and *DAPK1* expression is not unambiguous and may be associated with cell specialization. As for the mechanisms regulating the expression of these genes in HCV-infected hepatocarcinoma cells, they have not been studied yet.

p70/BCL6/TP53. *TP53* is a key activator of intrinsic apoptosis pathway. The NS3 (p70) protein affects *TP53* through interaction with the BCL6 (B-cell lymphoma 6 protein) TF. *TP53* is a HCC marker gene of and its low expression correlates with poor disease prognosis (Liu et al., 2012; Ye et al., 2017). BCL6 represses the *TP53* gene in lymphoid cells, and its constitutive expression protects B lymphocytes from DNA damage-induced apoptosis (Phan, Dalla-Favera, 2004; Jardin et al., 2007). However, the data describing the effect of HCV has on *TP53* and *BCL6* expression in these cells are associated with a possible mutation induction and are mutually exclusive (Machida et al., 2004; Tucci et al., 2013). The interaction of NS3 and BCL6 is discussed in (Han et al., 2016), but the specific mechanisms NS3 affects the TF activity have not been studied yet.

Discussion

The studied set of hypermethylated HCC marker genes (see Table 1) included 30 downexpressed and two over-expressed genes. Using ANDSystem, the regulatory pathways by which HCV proteins are able to influence the expression of these marker genes were reconstructed. The relationship between the pathways and HCC-associated key biological processes is shown in Figure 3. According to the published data, the *WT1*, *RUNX3*, *TP53*, and *SOCS1* genes are closely associated with apoptosis (Mayo et al., 1999; Loeb, 2006; Kim et al., 2019), while the *MGMT*, *TERT*, *RASSF1A*, and *WIF1* genes – with apoptosis and cell proliferation (He et al., 2005; Sarin et al., 2005; Choi et al., 2008; Feng et al., 2014; Chen J. et al., 2020; Ni et al., 2020).

The analysis showed the identified pathways could potentially be a part of the mechanism HCV proteins affect the activity of HCC marker genes. However, the effects the proteins have on the function of human regulatory proteins during PPI formation are currently poorly understood. This fact prevents us from unambiguous interpretation of the reconstructed pathways, because what determines if a regulatory pathway functions as an activator or suppressor of target gene expression is whether or not the ability to regulate gene expression remains in the host regulatory protein after its interaction with the viral protein. Investigation of these effects requires further experimental studies and computer molecular modeling.

The literature describes the effects viral proteins can have on the function of host proteins, e. g., the NS5A protein is known to bind to SMYD3 in the cytoplasm and inhibit SMYD3 translocation to the nucleus (Chen M. et al., 2016). If the regulatory proteins of a host organism are assumed to lose their ability to regulate gene expression due to the complexes formed with viral proteins, then the following effects can be expected: when considering the pathways regulating onco-suppressor expression, four of the seven pathways inhibiting *RASFI*, *RUNX3*, *WIF1*, and *DAPK1* will be suppressed, which can lead to their activation by HCV proteins, and that, in turn, will prevent carcinogenesis. In the remaining three pathways, the activation of *MGMT*, *SOCS1* and *TP53* will be suppressed, which may possibly have a protumor effect. In the presented pathways (see Fig. 2), *WT1*, *MGMT*, *SOCS1* and *TP53* are activated by the corresponding factors (Phan et al., 2004; Kohsaka et al., 2012; Yang C. et al., 2015; Zong et al., 2017), while the expression of *RASFI*, *RUNX3*, *WIF1*, and *DAPK1* genes is negatively controlled (Guo et al., 2011; Benderska, Schneider-Stock, 2014; Nowyhed et al., 2015; Lin et al., 2018).

In the case of the *TERT* and *WT1* genes that can be attributed to those with protumor activity, suppression of the *WT1* gene can be expected and will lead to a negative effect on carcinogenesis. As for the *TERT* gene involved in apoptosis suppression and stimulating cell proliferation (Nault et al., 2019; In der Stroth et al., 2020), according to our results (see Fig. 2), this gene was controlled through three different regulatory pathways. Its expression was activated by two pathways involving the STAT3 and NOTC1 host genes (Konnikova et al., 2005; Sawangarun et al., 2018), and one of the pathways suppressed the expression involving c-JUN/AP-1 (Takakura et al., 2005). The interaction with these host proteins could lead to a blockage of these regulatory pathways. The pathway involving c-JUN/AP-1 may be of particular interest in this

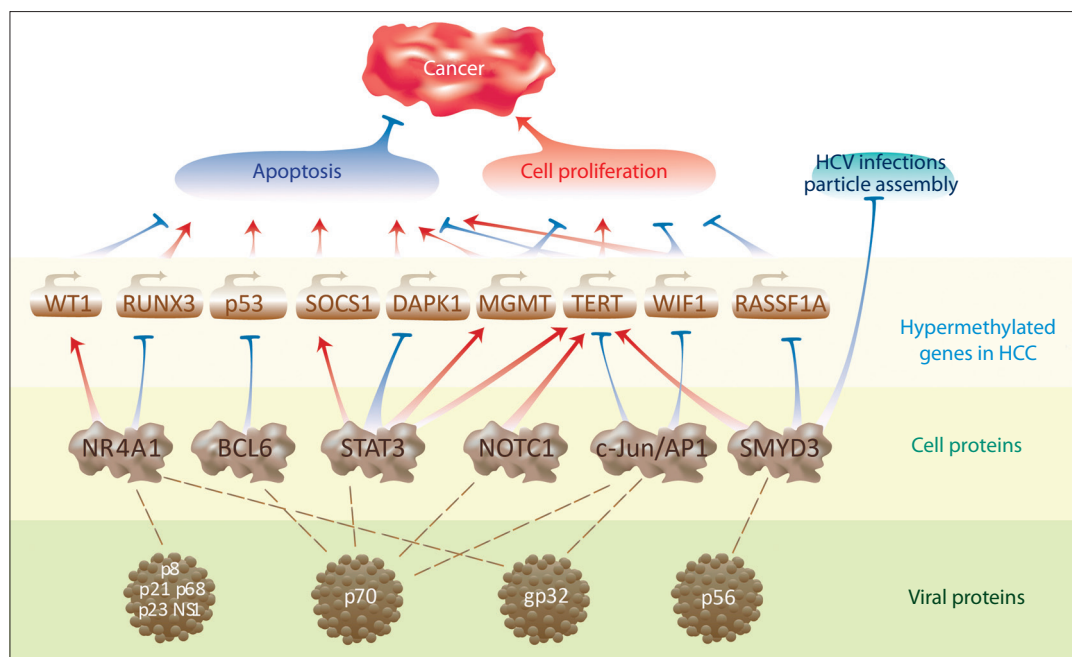


Fig. 3. Interrelation of the reconstructed regulatory pathways and the key biological processes associated with HCC.

respect, since inhibition of this TF by viral proteins (gp32 and NS3) could promote *TERT* activation. These assumptions are in good agreement with the data on differential gene expression in acute HCV infection when the infected cells showed an increased *TERT* expression (Papic et al., 2012), so this pathway can be a promising pharmacological target.

Thus, the considered assumptions lead one to conclude the presence of multidirectional regulation of the observed expression of HCC marker genes. This may indicate that not all regulatory pathways controlled by viral proteins can be attributed to HCC risk factors. However, the regulatory pathways that ensure the protumor activity of virus proteins undoubtedly deserve additional study to understand the mechanisms of virus-induced HCC carcinogenesis. In particular, the suppression of tumor suppressor gene expression by viral proteins can enhance the effect of their methylation in HCC or mimic this effect when these genes are not methylated, which can either provoke HCC onset or complicate its course.

Conclusion

Using the computer methods for gene network reconstruction available in the ANDSystem package, the statistically significant pathways for HCV proteins to regulate HCC gene markers have been established. The obtained results describe the potential mechanisms of the proteins involvement in HCC pathogenesis and may be useful for planning experimental studies to search for new targets for the development of drugs and prophylactic agents to reduce the risk of HCC developing in presence of HCV infection.

References

- Benderska N., Schneider-Stock R. Transcription control of DAPK. *Apoptosis*. 2014;19(2):298-305. DOI 10.1007/s10495-013-0931-6.
- Bragina E.Y., Tiys E.S., Rudko A.A., Ivanisenko V.A., Freidin M.B. Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks. *Infect. Genet. Evol.* 2016;46:118-123. DOI 10.1016/j.meegid.2016.10.030.
- Chen J., Li Z., Chen J., Du Y., Song W., Xuan Z., Zhao L., Song G., Song P., Zheng S. Downregulation of MGMT promotes proliferation of intrahepatic cholangiocarcinoma by regulating p21. *Clin. Transl. Oncol.* 2020;22(3):392-400. DOI 10.1007/s12094-019-02140-9.
- Chen M., Gan X., Yoshino K., Kitakawa M., Shoji I., Deng L., Hotta H. Hepatitis C virus NS5A protein interacts with lysine methyltransferase SET and MYND domain-containing 3 and induces activator protein 1 activation. *Microbiol. Immunol.* 2016;60:407-417. DOI 10.1111/1348-0421.12383.
- Chen P., Meng C., Wang Q., Yang X., Huang Z., Xing X., Lin Y., Liu X., Peng J., Lin Y. Death-associated protein kinase 1 suppresses hepatocellular carcinoma cell migration and invasion by upregulation of DEAD-box helicase 20. *Cancer Sci.* 2020;111(8):2803-2813. DOI 10.1111/cas.14499.
- Cheng J., Wei D., Ji Y., Chen L., Yang L., Li G., Wu L., Hou T., Xie L., Ding G., Li H., Li Y. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. *Genome Med.* 2018;10(1):42. DOI 10.1186/s13073-018-0548-z.
- Choi J., Southworth L.K., Sarin K.Y., Venteicher A.S., Ma W., Chang W., Cheung P., Jun S., Artandi M.K., Shah N., Kim S.K., Artandi S.E. TERT promotes epithelial proliferation through transcriptional control of a Myc- and Wnt-related developmental program. *PLoS Genet.* 2008;4(1):e10. DOI 10.1371/journal.pgen.0040010.
- de Chasse B., Navratil V., Tafforeau L., Hiet M.S., Aublin-Gex A., Agaugué S., Meiffren G., Pradezynski F., Faria B.F., Chantier T., Le Breton M., Pellet J., Davoust N., Mangeot P.E., Chaboud A., Penin F., Jacob Y., Vidalain P.O., Vidal M., André P., Rabourdin-Combe C., Lotteau V. Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 2008;4:230. DOI 10.1038/msb.2008.66.
- Deng Y., Yu B., Cheng Q., Jin J., You H., Ke R., Tang N., Shen Q., Shu H., Yao G., Zhang Z., Qin W. Epigenetic silencing of WIF-1 in hepatocellular carcinomas. *J. Cancer Res. Clin. Oncol.* 2010;136(8):1161-1167. DOI 10.1007/s00432-010-0763-5.
- D'souza S., Lau K.C., Coffin C.S., Patel T.R. Molecular mechanisms of viral hepatitis induced hepatocellular carcinoma. *World J. Gastroenterol.* 2020;26(38):5759-5783. DOI 10.3748/wjg.v26.i38.5759.

- Feng L., Li J., Yan L.D., Tang J. RASSF1A suppresses proliferation of cervical cancer cells. *Asian Pac. J. Cancer Prev.* 2014;15(14):5917-5920. DOI 10.7314/apjcp.2014.15.14.5917.
- Fernández-Barrena M.G., Arechederra M., Colyn L., Berasain C., Avila M.A. Epigenetics in hepatocellular carcinoma development and therapy: the tip of the iceberg. *JHEP Rep.* 2020;2(6):100167. DOI 10.1016/j.jhepr.2020.100167.
- Glotov A.S., Tiys E.S., Vashukova E.S., Pakin V.S., Demenkov P.S., Saik O.V., Ivanisenko T.V., Arzhanova O.N., Mozgovaya E.V., Zainulina M.S., Kolchanov N.A., Baranov V.S., Ivanisenko V.A. Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia asociome). *BMC Syst. Biol.* 2015;9(Suppl.2):S4. DOI 10.1186/1752-0509-9-S2-S4.
- Gui Y., Yeganeh M., Ramanathan S., Leblanc C., Pomerleau V., Ferbeyre G., Saucier C., Ilangumaran S. SOCS1 controls liver regeneration by regulating HGF signaling in hepatocytes. *J. Hepatol.* 2011;55(6):1300-1308. DOI 10.1016/j.jhep.2011.03.027.
- Guo N., Chen R., Li Z., Liu Y., Cheng D., Zhou Q., Zhou J., Lin Q. Hepatitis C virus core upregulates the methylation status of the RASSF1A promoter through regulation of SMYD3 in hilar cholangiocarcinoma cells. *Acta Biochim. Biophys. Sin. (Shanghai).* 2011;43(5):354-361. DOI 10.1093/abbs/gmr021.
- Han Y., Niu J., Wang D., Li Y. Hepatitis C virus protein interaction network analysis based on hepatocellular carcinoma. *PLoS One.* 2016;11(4):e0153882. DOI 10.1371/journal.pone.0153882.
- Hassan M., Ghozlan H., Abdel-Kader O. Activation of c-Jun NH2-terminal kinase (JNK) signaling pathway is essential for the stimulation of hepatitis C virus (HCV) non-structural protein 3 (NS3)-mediated cell growth. *Virology.* 2005;333(2):324-336. DOI 10.1016/j.virol.2005.01.008.
- Hassan M., Selimovic D., Ghozlan H., Abdel-Kader O. Induction of high-molecular-weight (HMW) tumor necrosis factor (TNF) alpha by hepatitis C virus (HCV) non-structural protein 3 (NS3) in liver cells is AP-1 and NF-κB-dependent activation. *Cell. Signal.* 2007;19(2):301-311. DOI 10.1016/j.cellsig.2006.07.002.
- He B., Reguart N., You L., Mazieres J., Xu Z., Lee A.Y., Mikami I., McCormick F., Jablons D.M. Blockade of Wnt-1 signaling induces apoptosis in human colorectal cancer cells containing downstream mutations. *Oncogene.* 2005;24(18):3054-3058. DOI 10.1038/sj.onc.1208511.
- Hernandez-Meza G., von Felden J., Gonzalez-Kozlova E.E., Garcia-Lezana T., Peix J., Portela A., Craig A.J., Sayols S., Schwartz M., Losic B., Mazzaferro V., Esteller M., Llovet J.M., Villanueva A. DNA methylation profiling of human hepatocarcinogenesis. *Hepatology.* 2021;74(1):183-199. DOI 10.1002/hep.31659.
- Hoshida Y., Villanueva A., Kobayashi M., Peix J., Chiang D.Y., Camargo A., Gupta S., Moore J., Wrobel M.J., Lerner J., Reich M., Chan J.A., Glickman J.N., Ikeda K., Hashimoto M., Watanabe G., Daidone M.G., Roayaie S., Schwartz M., Thung S., Salvesen H.B., Gabriel S., Mazzaferro V., Bruix J., Friedman S.L., Kumada H., Llovet J.M., Golub T.R. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N. Engl. J. Med.* 2008;359(19):1995-2004. DOI 10.1056/NEJMoa0804525.
- Huang L., Li M.X., Wang L., Li B.K., Chen G.H., He L.R., Xu L., Yuan Y.F. Prognostic value of Wnt inhibitory factor-1 expression in hepatocellular carcinoma that is independent of gene methylation. *Tumour Biol.* 2011;32(1):233-240. DOI 10.1007/s13277-010-0117-6.
- In der Stroth L., Tharehalli U., Günes C., Lechel A. Telomeres and telomerase in the development of liver cancer. *Cancers (Basel).* 2020;12(8):2048. DOI 10.3390/cancers12082048.
- Ivanisenko N.V., Seyrek K., Kolchanov N.A., Ivanisenko V.A., Lavrik I.N. The role of death domain proteins in host response upon SARS-CoV-2 infection: modulation of programmed cell death and translational applications. *Cell Death Discov.* 2020;6:101. DOI 10.1038/s41420-020-00331-w.
- 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 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., Cherezis S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Yu.I., Voevoda S.M., Karpenko T.N., Velichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolomics and gene regulatory networks analysis reveal the role of non-structural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci. Rep.* 2022;12:19977. DOI 10.1038/s41598-022-24170-0.
- 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.
- Iwai A., Takegami T., Shiozaki T., Miyazaki T. Hepatitis C virus NS3 protein can activate the Notch-signaling pathway through binding to a transcription factor, SRCAP. *PLoS One.* 2011;6(6):e20718. DOI 10.1371/journal.pone.0020718.
- Jardin F., Ruminy P., Bastard C., Tilly H. The BCL6 proto-oncogene: a leading role during germinal center development and lymphomagenesis. *Pathol. Biol. (Paris).* 2007;55(1):73-83. DOI 10.1016/j.patbio.2006.04.001.
- Jaroszewicz J., Flisiak-Jackiewicz M., Lebensztejn D., Flisiak R. Current drugs in early development for treating hepatitis C virus-related hepatic fibrosis. *Expert Opin. Investig. Drugs.* 2015;24(9):1229-1239. DOI 10.1517/13543784.2015.1057568.
- Jiang L.H., Hao Y.L., Zhu J.W. Expression and prognostic value of HER-2/neu, STAT3 and SOCS3 in hepatocellular carcinoma. *Clin. Res. Hepatol. Gastroenterol.* 2019;43(3):282-291. DOI 10.1016/j.clinre.2018.09.011.
- Jing W., Peng R., Li X., Lv S., Duan Y., Jiang S. Study on the prognostic values of TTC36 correlated with immune infiltrates and its methylation in hepatocellular carcinoma. *J. Immunol. Res.* 2022;2022:7267131. DOI 10.1155/2022/7267131.
- Kim B.R., Park S.H., Jeong Y.A., Na Y.J., Kim J.L., Jo M.J., Jeong S., Yun H.K., Oh S.C., Lee D.H. RUNX3 enhances TRAIL-induced apoptosis by upregulating DR5 in colorectal cancer. *Oncogene.* 2019;38:3903-3918. DOI 10.1038/s41388-019-0693-x.
- Kohsaka S., Wang L., Yachi K., Mahabir R., Narita T., Itoh T., Tanino M., Kimura T., Nishihara H., Tanaka S. STAT3 inhibition overcomes temozolomide resistance in glioblastoma by downregulating MGMT expression. *Mol. Cancer Ther.* 2012;11(6):1289-1299. DOI 10.1158/1535-7163.MCT-11-0801.
- Konnikova L., Simeone M.C., Kruger M.M., Kotecki M., Cochran B.H. Signal transducer and activator of transcription 3 (STAT3) regulates human telomerase reverse transcriptase (hTERT) expression in human cancer and primary cells. *Cancer Res.* 2005;65(15):6516-6520. DOI 10.1158/0008-5472.CAN-05-0924.
- Li B., Li X., Li Y., Guo H., Sun S.Y., He Q.Q., Wang Y., Luo J., Wen J.F., Zheng H., Feng D.Y. The effects of hepatitis C virus non-structural protein 3 on cell growth mediated by extracellular signal-related kinase cascades in human hepatocytes in vitro. *Int. J. Mol. Med.* 2010;26(2):273-279. DOI 10.3892/ijmm.00000462.
- Lin B., Hong H., Jiang X., Li C., Zhu S., Tang N., Wang X., She F., Chen Y. c-Jun suppresses the expression of WNT inhibitory factor 1 through transcriptional regulation and interaction with DNA methyltransferase 1 in gallbladder cancer. *Mol. Med. Rep.* 2018;17(6):8180-8188. DOI 10.3892/mmr.2018.8890.
- Liu J., Ma Q., Zhang M., Wang X., Zhang D., Li W., Wang F., Wu E. Alterations of TP53 are associated with a poor out-come for patients

- with hepatocellular carcinoma: evidence from a systematic review and meta-analysis. *Eur. J. Cancer*. 2012;48(15):2328-2338. DOI 10.1016/j.ejca.2012.03.001.
- Llovet J.M., Zucman-Rossi J., Pikarsky E., Sangro B., Schwartz M., Sherman M., Gores G. Hepatocellular carcinoma. *Nat. Rev. Dis. Primers*. 2016;2:16018. DOI 10.1038/nrdp.2016.18.
- Loeb D.M. WT1 influences apoptosis through transcriptional regulation of Bcl-2 family members. *Cell Cycle*. 2006;5(12):1249-1253. DOI 10.4161/cc.5.12.2807.
- 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.
- Machida K., Cheng K.T., Sung V.M., Shimodaira S., Lindsay K.L., Levine A.M., Lai M.Y., Lai M.M. Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and proto-oncogenes. *Proc. Natl. Acad. Sci. USA*. 2004;101(12):4262-4267. DOI 10.1073/pnas.0303971101.
- Mayo M.W., Wang C.Y., Drouin S.S., Madrid L.V., Marshall A.F., Reed J.C., Weissman B.E., Baldwin A.S. WT1 modulates apoptosis by transcriptionally upregulating the *bcl-2* proto-oncogene. *EMBO J*. 1999;18(14):3990-4003. DOI 10.1093/emboj/18.14.3990.
- McGlynn K.A., Petrick J.L., El-Serag H.B. Epidemiology of hepatocellular carcinoma. *Hepatology*. 2021;73(Suppl.1):4-13. DOI 10.1002/hep.31288.
- Mžik M., Chmelařová M., John S., Laco J., Slabý O., Kiss I., Bohovicová L., Palička V., Nekvindová J. Aberrant methylation of tumour suppressor genes *WT1*, *GATA5* and *PAX5* in hepatocellular carcinoma. *Clin. Chem. Lab. Med.* 2016;54(12):1971-1980. DOI 10.1515/cebm-2015-1198.
- Nault J.C., Ningharhari M., Rebouissou S., Zucman-Rossi J. The role of telomeres and telomerase in cirrhosis and liver cancer. *Nat. Rev. Gastroenterol. Hepatol*. 2019;16(9):544-558. DOI 10.1038/s41575-019-0165-3.
- Neumann O., Kesselmeier M., Geffers R., Pellegrino R., Radlwimmer B., Hoffmann K., Ehemann V., Schemmer P., Schirmacher P., Lorenzo Bermejo J., Longerich T. Methylation analysis and integrative profiling of human HCCs identify novel protumorigenic factors. *Hepatology*. 2012;56(5):1817-1827. DOI 10.1002/hep.25870.
- Ni Y., Gu J., Wu J., Xu L., Rui Y. MGMT-mediated neuron apoptosis in injured rat spinal cord. *Tissue Cell*. 2020;62:101311. DOI 10.1016/j.tice.2019.101311.
- Nowyhed H.N., Huynh T.R., Blatchley A., Wu R., Thomas G.D., Hedrick C.C. The nuclear receptor Nr4a1 controls CD8 T cell development through transcriptional suppression of Runx3. *Sci. Rep*. 2015;5:9059. DOI 10.1038/srep09059.
- Oversoe S.K., Clement M.S., Pedersen M.H., Weber B., Aagaard N.K., Villadsen G.E., Grønbaek H., Hamilton-Dutoit S.J., Sorensen B.S., Kelsen J. TERT promoter mutated circulating tumor DNA as a biomarker for prognosis in hepatocellular carcinoma. *Scand. J. Gastroenterol*. 2020;55(12):1433-1440. DOI 10.1080/00365521.2020.1837928.
- Papic N., Maxwell C.I., Delker D.A., Liu S., Heale B.S., Hagedorn C.H. RNA-sequencing analysis of 5' capped RNAs identifies many new differentially expressed genes in acute hepatitis C virus infection. *Viruses*. 2012;4:581-612. DOI 10.3390/v4040581.
- Phan R.T., Dalla-Favera R. The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature*. 2004;432(7017):635-639. DOI 10.1038/nature03147.
- Rabaan A.A., Al-Ahmed S.H., Bazzi A.M., Alfouzan W.A., Alsuliman S.A., Aldrazi F.A., Haque S. Overview of hepatitis C infection, molecular biology, and new treatment. *J. Infect. Public Health*. 2020;13(5):773-783. DOI 10.1016/j.jiph.2019.11.015.
- Revill K., Wang T., Lachenmayer A., Kojima K., Harrington A., Li J., Hoshida Y., Llovet J.M., Powers S. Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology*. 2013;145(6):1424-1435.e1-25. DOI 10.1053/j.gastro.2013.08.055.
- Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofstaedt R., Lavrik I.N., Rogaev E.I., Ivanisenko V.A. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018;11(Suppl.1):15. DOI 10.1186/s12920-018-0331-4.
- Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interaction of the hepatitis C virus: literature mining with ANDSystem. *Virus Res*. 2016;218:40-48. DOI 10.1016/j.virusres.2015.12.003.
- Sarin K.Y., Cheung P., Gilson D., Lee E., Tennen R.I., Wang E., Artandi M.K., Oro A.E., Artandi S.E. Conditional telomerase induction causes proliferation of hair follicle stem cells. *Nature*. 2005;436(7053):1048-1052. DOI 10.1038/nature03836.
- Sawangarun W., Mandasari M., Aida J., Morita K.I., Kayamori K., Ikeda T., Sakamoto K. Loss of Notch1 predisposes oro-esophageal epithelium to tumorigenesis. *Exp. Cell Res*. 2018;372(2):129-140. DOI 10.1016/j.yexcr.2018.09.019.
- Schulze K., Imbeaud S., Letouzé E., Alexandrov L.B., Calderaro J., Rebouissou S., Couchy G., Meiller C., Shinde J., Soysouvanh F., Calatayud A.L., Pinyol R., Pelletier L., Balabaud C., Laurent A., Blanc J.F., Mazzaferro V., Calvo F., Villanueva A., Nault J.C., Bioulac-Sage P., Stratton M.R., Llovet J.M., Zucman-Rossi J. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet*. 2015;47(5):505-511. DOI 10.1038/ng.3252.
- Sera T., Hiasa Y., Mashiba T., Tokumoto Y., Hirooka M., Konishi I., Matsuura B., Michitaka K., Uda K., Onji M. Wilms' tumour 1 gene expression is increased in hepatocellular carcinoma and associated with poor prognosis. *Eur. J. Cancer*. 2008;44(4):600-608. DOI 10.1016/j.ejca.2008.01.008.
- Song Z., Li Z., Han W., Zhu C., Lou N., Li X., Luo G., Peng S., Li G., Zhao Y., Guo Y. Low DAPK1 expression correlates with poor prognosis and sunitinib resistance in clear cell renal cell carcinoma. *Aging (Albany NY)*. 2020;13(2):1842-1858. DOI 10.18632/aging.103638.
- Takakura M., Kyo S., Inoue M., Wright W.E., Shay J.W. Function of AP-1 in transcription of the telomerase reverse transcriptase gene (*TERT*) in human and mouse cells. *Mol. Cell Biol*. 2005;25(18):8037-8043. DOI 10.1128/MCB.25.18.8037-8043.2005.
- 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.
- Tucci F.A., Broering R., Johansson P., Schlaak J.F., Küppers R. B cells in chronically hepatitis C virus-infected individuals lack a virus-induced mutation signature in the *TP53*, *CTNBN1*, and *BCL6* genes. *J. Virol*. 2013;87(5):2956-2962. DOI 10.1128/JVI.03081-12.
- Xu R.H., Wei W., Krawczyk M., Wang W., Luo H., Flagg K., Yi S., Shi W., Quan Q., Li K., Zheng L., Zhang H., Caughey B.A., Zhao Q., Hou J., Zhang R., Xu Y., Cai H., Li G., Hou R., Zhong Z., Lin D., Fu X., Zhu J., Duan Y., Yu M., Ying B., Zhang W., Wang J., Zhang E., Zhang C., Li O., Guo R., Carter H., Zhu J.K., Hao X., Zhang K. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat. Mater*. 2017;16(11):1155-1161. DOI 10.1038/nmat4997.
- Yang C., Zhang Y., Wang J., Li L., Wang L., Hu M., Xu M., Long Y., Rong R., Zhu T. A novel cyclic helix B peptide inhibits dendritic cell maturation during amelioration of acute kidney graft rejection through Jak-2/STAT3/SOCS1. *Cell Death Dis*. 2015;6(11):e1993. DOI 10.1038/cddis.2015.338.
- Ye S., Zhao X.Y., Hu X.G., Li T., Xu Q.R., Yang H.M., Huang D.S., Yang L. TP53 and RET may serve as biomarkers of prognostic evaluation and targeted therapy in hepatocellular carcinoma. *Oncol. Rep*. 2017;37(4):2215-2226. DOI 10.3892/or.2017.5494.
- Zhang C., Li J., Huang T., Duan S., Dai D., Jiang D., Sui X., Li D., Chen Y., Ding F., Huang C., Chen G., Wang K. Meta-analysis of

- DNA methylation biomarkers in hepatocellular carcinoma. *Oncotarget*. 2016;7(49):81255-81267. DOI 10.18632/oncotarget.13221.
- Zhang H., Weng X., Ye J., He L., Zhou D., Liu Y. Promoter hypermethylation of *TERT* is associated with hepatocellular carcinoma in the Han Chinese population. *Clin. Res. Hepatol. Gastroenterol.* 2015;39(5):600-609. DOI 10.1016/j.clinre.2015.01.002.
- Zhu Z., Tran H., Mathahs M.M., Moninger T.O., Schmidt W.N. HCV induces telomerase reverse transcriptase, increases its catalytic activity, and promotes caspase degradation in infected human hepatocytes. *PLoS One*. 2017;12(1):e0166853. DOI 10.1371/journal.pone.0166853.
- Zhu Z., Wilson A.T., Gopalakrishna K., Brown K.E., Luxon B.A., Schmidt W.N. Hepatitis C virus core protein enhances Telomerase activity in Huh7 cells. *J. Med. Virol.* 2010;82(2):239-248. DOI 10.1002/jmv.21644.
- Zong C., Qin D., Yu C., Gao P., Chen J., Lu S., Zhang Y., Liu Y., Yang Y., Pu Z., Li X., Fu Y., Guan Q., Wang X. The stress-response molecule NR4A1 resists ROS-induced pancreatic β -cells apoptosis via WT1. *Cell Signal*. 2017;35:129-139. DOI 10.1016/j.cellsig.2017.03.012.
- Zucman-Rossi J., Villanueva A., Nault J.C., Llovet J.M. Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology*. 2015;149(5):1226-1239.e4. DOI 10.1053/j.gastro.2015.05.061.

ORCID ID

E.A. Antropova orcid.org/0000-0003-2158-3252
P.S. Demenkov orcid.org/0000-0001-9433-8341

Acknowledgements. The present study was supported by project No. 075-15-2021-944 of the Ministry of Science and Higher Education of the Russian Federation as a part of ERA-NET Target Identification and Drug Development in Liver Cancer (TAIGA).

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

Received November 15, 2022. Revised November 22, 2022. Accepted November 22, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Rational metabolic engineering of *Corynebacterium glutamicum* to create a producer of L-valine

M.E. Sheremetieva¹✉, K.E. Anufriev¹, T.M. Khlebodarova^{2, 3}, N.A. Kolchanov^{2, 3}, A.S. Yanenko¹

¹ NRC "Kurchatov Institute", Kurchatov Genomic Center, Moscow, Russia

² Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

✉ m.e.sheremetieva@gmail.com

Abstract. L-Valine is one of the nine amino acids that cannot be synthesized *de novo* by higher organisms and must come from food. This amino acid not only serves as a building block for proteins, but also regulates protein and energy metabolism and participates in neurotransmission. L-Valine is used in the food and pharmaceutical industries, medicine and cosmetics, but primarily as an animal feed additive. Adding L-valine to feed, alone or mixed with other essential amino acids, allows for feeds with lower crude protein content, increases the quality and quantity of pig meat and broiler chicken meat, as well as improves reproductive functions of farm animals. Despite the fact that the market for L-valine is constantly growing, this amino acid is not yet produced in our country. In modern conditions, the creation of strains-producers and organization of L-valine production are especially relevant for Russia. One of the basic microorganisms most commonly used for the creation of amino acid producers, along with *Escherichia coli*, is the soil bacterium *Corynebacterium glutamicum*. This review is devoted to the analysis of the main strategies for the development of L-valine producers based on *C. glutamicum*. Various aspects of L-valine biosynthesis in *C. glutamicum* are reviewed: process biochemistry, stoichiometry and regulation, enzymes and their corresponding genes, export and import systems, and the relationship of L-valine biosynthesis with central cell metabolism. Key genetic elements for the creation of *C. glutamicum*-based strains-producers are identified. The use of metabolic engineering to enhance L-valine biosynthesis reactions and to reduce the formation of byproducts is described. The prospects for improving strains in terms of their productivity and technological characteristics are shown. The information presented in the review can be used in the production of producers of other amino acids with a branched side chain, namely L-leucine and L-isoleucine, as well as D-pantothenate.

Key words: *Corynebacterium glutamicum*; L-valine; metabolic engineering; producer strain.

For citation: Sheremetieva M.E., Anufriev K.E., Khlebodarova T.M., Kolchanov N.A., Yanenko A.S. Rational metabolic engineering of *Corynebacterium glutamicum* to create a producer of L-valine. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):743-757. DOI 10.18699/VJGB-22-90

Рациональная метаболическая инженерия *Corynebacterium glutamicum* для продукции L-валина

М.Е. Шереметьева¹✉, К.Э. Ануфриев¹, Т.М. Хлебодарова^{2, 3}, Н.А. Колчанов^{2, 3}, А.С. Яненко¹

¹ Национальный исследовательский центр «Курчатовский институт», Курчатовский геномный центр, Москва, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

✉ m.e.sheremetieva@gmail.com

Аннотация. L-Валин – одна из девяти аминокислот, которые не могут быть синтезированы *de novo* высшими организмами и должны поступать с пищей. Эта аминокислота не только служит строительным материалом для белков, но также регулирует белковый и энергетический обмен и участвует в нейротрансмиссии. L-Валин используется в пищевой и фармацевтической промышленности, медицине и косметике, но в первую очередь в качестве кормовой добавки для животных. Добавление L-валина в корм отдельно или в смеси с другими незаменимыми аминокислотами позволяет использовать корма с меньшим содержанием сырого белка, повышает качество и количество мяса свиней и цыплят-бройлеров, а также улучшает репродуктивные функции сельскохозяйственных животных. Несмотря на то что рынок L-валина постоянно растет, в нашей стране эта аминокислота пока не производится. В современных условиях создание штаммов-продуцентов и организация производства L-валина для России особенно актуальны. Один из наиболее часто используемых базовых микроорганизмов для создания продуцентов аминокислот наряду с *Escherichia coli* – почвенная бактерия *Corynebacterium glutamicum*. Обзор посвящен анализу основных стратегий разработки продуцентов L-валина на базе *C. glutamicum*. Рассмотрены различные аспекты биосинтеза L-валина у коринебактерий: биохимия, стехиометрия и регуляция процес-

са, ферменты и соответствующие им гены, системы экспорта и импорта, связь биосинтеза L-валина с центральным метаболизмом клетки. Выявлены ключевые генетические элементы для создания штаммов-продуцентов на основе *C. glutamicum*. Описано использование метаболической инженерии для усиления реакций биосинтеза L-валина и уменьшения образования побочных продуктов. Показаны перспективы усовершенствования штаммов с точки зрения повышения их продуктивности и улучшения технологических характеристик. Информация, представленная в обзоре, может быть использована при получении продуцентов других аминокислот с разветвленной боковой цепью – L-лейцина и L-изолейцина, а также D-пантотената.

Ключевые слова: *Corynebacterium glutamicum*; L-валин; метаболическая инженерия; штамм-продуцент.

Introduction

L-Valine is a proteinogenic branched-chain amino acid (BCAA), which also include L-leucine and L-isoleucine (hereinafter referred to as valine, leucine, isoleucine). These are essential amino acids that are not synthesized in humans or animals and must be present in the diet. Therefore, these amino acids are mainly used in the animal feed industry and as a dietary supplement for humans (Karau, Grayson, 2014). The former is particularly relevant to the global task of intensifying livestock production. Adding valine to feeds, either alone or mixed with other BCAAs, leads to improved meat quality and quantity in pigs and broiler chickens, increased egg production in chickens, increased lactation, milk fat content and appetite in pigs (Zheng et al., 2017; Che et al., 2021; Jian et al., 2021). A balance between different BCAAs, however, must be maintained, as its disruption can reduce the observed beneficial effects (Holen et al., 2022).

In addition to the livestock and food industries, BCAAs find their application in pharmacology and medicine. BCAAs not only serve as building blocks for proteins, but also participate in the regulation of protein and energy metabolism, their consumption increases exercise tolerance and accelerates fatty acid oxidation (Kainulainen et al., 2013). They are useful as supplements for chronic liver disease (Kawaguchi et al., 2011) and for stimulating macrophage phagocytosis of multidrug-resistant bacterial pathogens (Chen et al., 2017). As with feed additives, when using BCAAs for food and drug production their concentration should be chosen carefully. Excess BCAA in human plasma increases the risk of several diseases, including type 2 diabetes, metabolic syndrome, obesity, hypertension, and cardiovascular disease (Holeček, 2018; Dimou et al., 2022), but has little effect on athletes who are characterized by high physical activity (Shou et al., 2019).

Amino acids account for 62.3 % of the global feed supplement market, which is projected to be \$34.2 billion in 2022. L-lysine and L-methionine (hereinafter referred to as lysine, methionine) are the most in demand; the valine market is one of the fastest growing, along with L-threonine (hereinafter referred to as threonine) and L-tryptophan. Consumption of feed amino acids is concentrated in Europe, USA and China; Russia's share is less than 2 %, but shows a growing trend: from 2016 to 2017 the increase was 2.9 % (<https://agri-news.ru/zhurnal/2018/32018/ekonomika-menedzhment-ryinki/ryinok-kormovyix-aminokislot.html>). Currently, all valine on the Russian market is imported from China, one of the main producers of this amino acid.

Amino acids can be isolated from natural protein sources, obtained by chemical synthesis, as well as by a microbiological method based on the use of strain-producers. The latter option has important advantages: it allows to use renewable

raw material resources and to produce biologically active L-enantiomers of amino acids separately, rather than mixed with D-enantiomers, and is therefore used by leading valine producers (D'Este et al., 2017).

Amino acid producers are developed from *Escherichia coli* and *Corynebacterium glutamicum*. *E. coli* is a thoroughly studied bacterium for which an extensive toolkit of genetic modification is available. Due to that fact producer strains were previously derived mainly from *E. coli*. However, strains of *C. glutamicum* created by selection were also used. The history of their use for amino acid production goes back more than 60 years (Leuchtenberger et al., 2005). In recent decades, having made considerable progress both in understanding the metabolism of *C. glutamicum* and in improving methods for modifying their genome, developers of producer strains have increasingly begun to favor *Corynebacteria*.

Corynebacteria are nonpathogenic, GC-rich gram-positive bacteria, which, unlike *E. coli*, do not form endotoxins that cause allergic reactions in higher organisms. They are also characterized by flexible cellular metabolism, genetic stability, stress tolerance, including resistance to high concentrations of carbon sources and metabolites, and the ability to synthesize the target product when growth stops (Baritugo et al., 2018). Valine produced by fermentation using *C. glutamicum* strains is now recognized as safe (non-toxic and non-carcinogenic) for use as a food and feed additive and for other biological purposes (Kang et al., 2020).

This review presents the main strategies for increasing valine production by *C. glutamicum* cells. It also summarizes the achievements in the creation of valine-producing strains. In addition to obtaining valine, some aspects of obtaining isoleucine, leucine, and D-pantothenate (hereinafter, pantothenate) are also discussed because the biosynthesis of these compounds involves the same metabolic precursors, cofactors, and enzymes as does valine biosynthesis.

Valine biosynthesis in *C. glutamicum* and mechanisms of regulation of this process

Valine (2-amino-3-methylbutyric acid) is synthesized from pyruvate (pyruvic acid) through four consecutive reactions involving (Fig. 1): 1) condensation of two pyruvate molecules to form acetolactate, catalyzed by acetolactate synthase (AHAS); 2) NADPH-dependent conversion of acetolactate to 2,3-dihydroxyketoisovalerate, catalyzed by acetolactate reductoisomerase (AHAIR); 3) conversion of 2,3-dihydroxyketoisovalerate to 2-ketoisovalerate catalyzed by dihydroxyacid dehydratase (DHAD); 4) NADPH-dependent formation of valine from 2-ketoisovalerate catalyzed by BCAA transaminase (BCAT) and several other transaminases (Yamamoto et al., 2017).

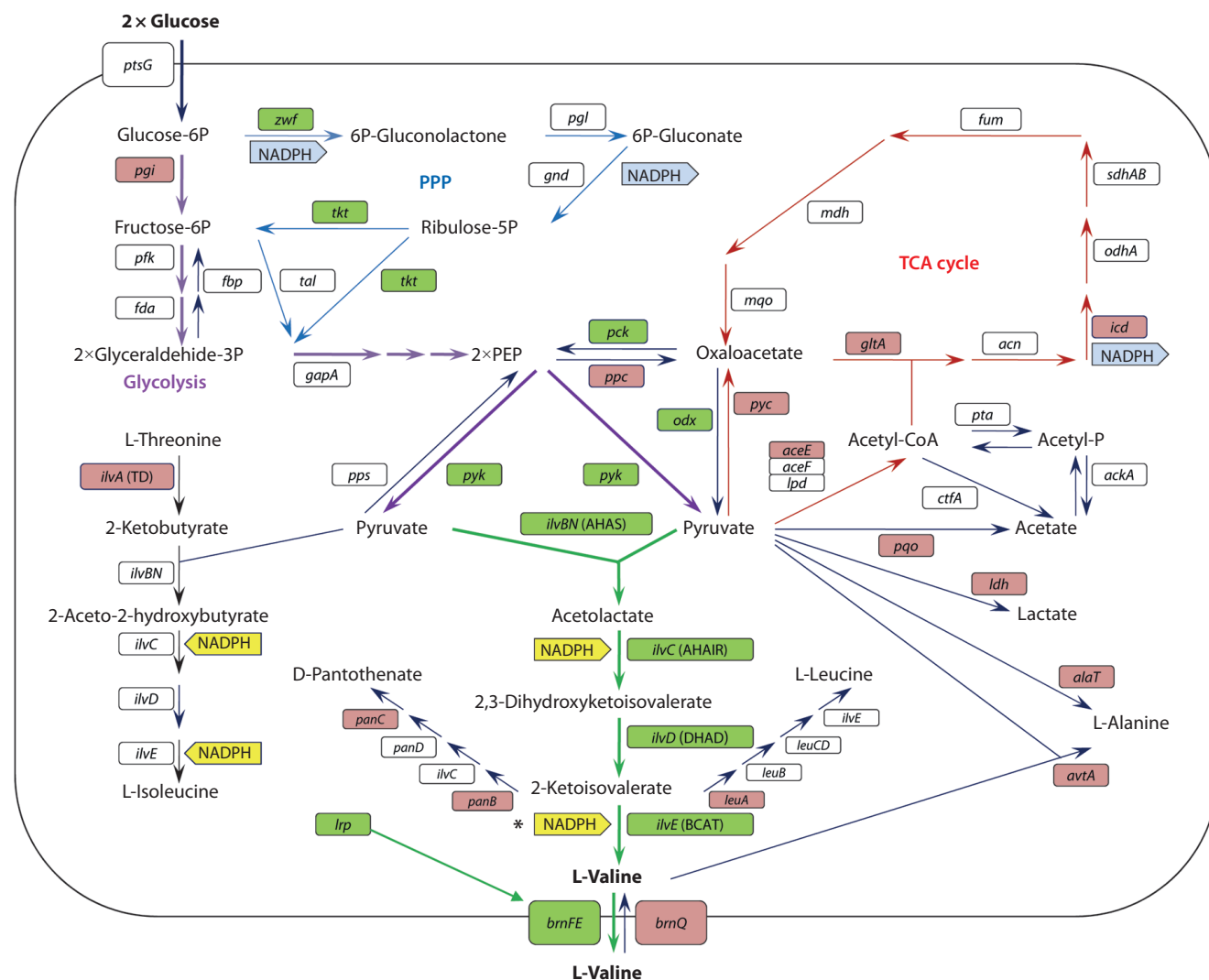


Fig. 1. Biosynthesis of valine and related metabolic pathways in *C. glutamicum* cells.

The genes whose increased expression leads to an increase (green) or decrease (red) in valine production are highlighted. A detailed description and transcript of the abbreviations are given in the text. An asterisk marks the reaction in which NADPH is used indirectly (see explanations in the text).

During synthesis, 2 mol of pyruvate and 2 mol of reducing equivalents in the form of reduced nicotinamide dinucleotide phosphate (NADPH) are consumed to produce 1 mol of valine. Pyruvate is formed from phosphoenolpyruvate (PEP) in glycolysis, which converts 1 mol of glucose to 2 mol of pyruvate. The main source of NADPH in *Corynebacteria* is the pentose-phosphate pathway (PPP) (Marx et al., 1997).

2-Ketoisovalerate is also a precursor of leucine and pantothenate (Park, Lee, 2010). In most microorganisms, including *C. glutamicum*, the same four enzymes catalyze isoleucine biosynthesis from pyruvate and 2-ketobutyrate. The latter is formed from threonine by threonine dehydratase (TD). Thus, the processes of biosynthesis of all three BCAAs (valine, leucine, and isoleucine) are closely linked. The synthesized BCAAs are removed from the cell by one export system, BrnFE (Lange et al., 2012).

A schematic of valine biosynthesis and related metabolic pathways in *C. glutamicum* is shown in Fig. 1. The key enzyme in the biosynthesis pathway of valine and other BCAAs is acetolactate synthase AHAS, which catalyzes the formation

of either acetolactate from two pyruvate molecules (in valine and leucine biosynthesis) or 2-aceto-2-hydroxybutyrate from pyruvate and 2-ketobutyrate (in isoleucine biosynthesis). In contrast to *E. coli*, only one form of the AHAS enzyme was found in *C. glutamicum* (Keilhauer et al., 1993), a tetramer consisting of two catalytic and two regulatory subunits (Liu et al., 2016). The catalytic and regulatory subunits of AHAS are encoded by the *ilvB* and *ilvN* genes, respectively. Together with the *ilvC* gene encoding the acetolactate reductoisomerase AHAI, these two genes form the operon *ilvBNC* with two additional promoters within it. Expression from the three promoters leads to the formation of transcripts of different lengths (Fig. 2). The *ilvC* gene is transcribed as part of all mRNAs; its expression efficiency is the highest among the three genes (Keilhauer et al., 1993; Morbach et al., 2000).

The expression of the operon *ilvBNC* is thought to be controlled by the mechanism of transcription attenuation, which is realized through the formation of secondary RNA structures (hairpins) on the transcript, i.e., transcription terminators that arise in the regulatory region in the presence of high

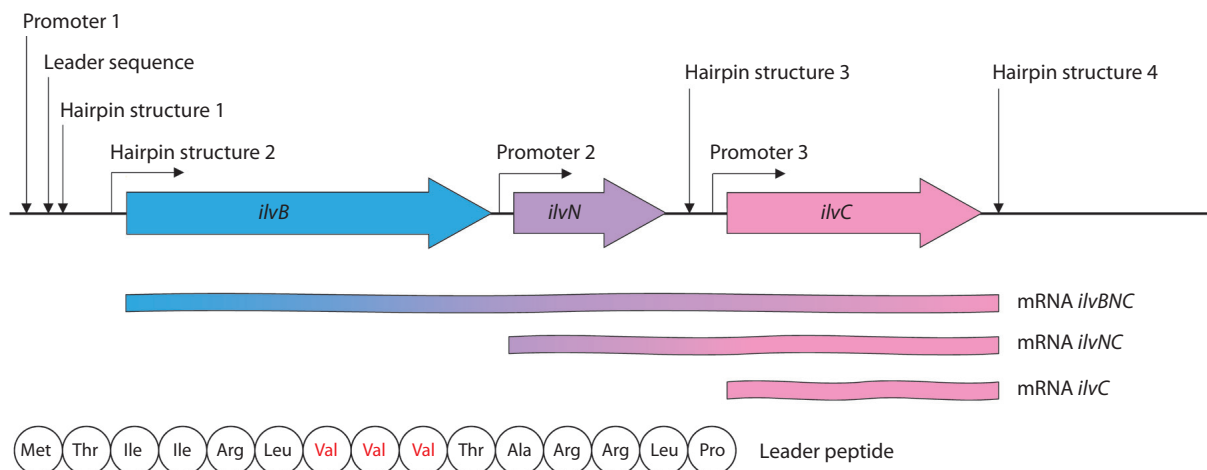


Fig. 2. Organization of the *ilvBNC* operon in *C. glutamicum* and regulation of its expression (adapted from the review (Wang et al., 2018)).

concentrations of BCAA (see Fig. 2). The regulatory region is upstream of the *ilvB* gene; in addition to the sites responsible for hairpin formation, it also encodes a leader peptide (25 amino acids) enriched with isoleucine (2), valine (3), and leucine (2) residues. It is assumed that this peptide is a sensor element of the regulatory system: when the concentration of any of the BCAAs in the cells is low, its translation is slowed down, resulting in no formation of terminator hairpin.

When one or more BCAA were lacking, the expression of operon *ilvBNC* doubled; replacement of valine residues in the leader peptide with alanine residues led to loss of valine effect on expression (Morbach et al., 2000). A significant increase in the expression of the operon *ilvBNC* in the presence of 2-ketobutyrate was observed (Eggeling et al., 1987; Keilhauer et al., 1993; Morbach et al., 2000). The mechanism of this regulation has not been investigated.

AHAS activity is strictly inhibited by valine ($K_i = 0.9$ mM) as well as leucine ($K_i = 6.0$ mM) and isoleucine ($K_i = 3.1$ mM) by a feedback mechanism through amino acid attachment to the regulatory subunit of the enzyme (Eggeling et al., 1987; Morbach et al., 2000; Leyval et al., 2003; Elišáková et al., 2005) and is also competitively inhibited by 2-ketoisovalerate (Krause et al., 2010a). Regardless of the number of BCAAs present (one, two, or all three), the degree of inhibition of AHAS activity does not exceed 57 % (Elišáková et al., 2005).

It should be noted that AHAS has lower substrate specificity towards pyruvate ($K_m = 8.3$ mM) (Leyval et al., 2003) than towards 2-ketobutyrate ($K_m = 4.8$ mM) (Eggeling et al., 1987), therefore, all other conditions being equal, the reaction of pyruvate condensation with 2-ketobutyrate leading to isoleucine synthesis is preferred.

As for AHAIIR (product of the *ilvC* gene), which catalyzes the isomerization step and the conversion of acetolactate to 2,3-dihydroxyketoisovalerate and 2-aceto-2-hydroxybutyrate to 2,3-dihydroxy-3-methylvalerate in the isoleucine synthesis pathway, its activity depends on the presence of NADPH and is inhibited by the feedback mechanism of valine and leucine, but not isoleucine (Leyval et al., 2003; Lee et al., 2019).

There is little information about the regulation of the activity of the enzymes controlling the third and fourth, final,

steps of valine synthesis in *C. glutamicum*, as well as about the regulation of the genes encoding their structure. It is only known that the activity of dihydroxyacid dehydratase DHAD (product of the *ilvD* gene) is weakly inhibited by valine and leucine and not inhibited by isoleucine (Leyval et al., 2003), and the activity of transaminase BCAT (product of the *ilvE* gene) depends on NADPH availability. The donor amino group in the transamination reaction is L-glutamate (hereafter referred to as glutamate), which is converted to 2-ketoglutarate; NADPH is required for glutamate regeneration by glutamate dehydrogenase. It has been shown that the reaction catalyzed by glutamate dehydrogenase is the main reaction of nitrogen assimilation under conditions of ammonia excess, which usually take place in amino acid production processes (Burkovski, 2003).

It has also been shown that alanine/valintransaminase (a product of the *avtA* gene) is involved in valine biosynthesis. Alanine/valintransaminase uses L-alanine (hereafter referred to as alanine) or α -aminobutyrate as an amino group donor instead of glutamate (Leyval et al., 2003).

Analysis of the dynamics of changes in the concentrations of the metabolites of valine biosynthesis using a kinetic model in *C. glutamicum* strain ATCC 13032 Δ *panBC* Δ *ilvA* pJC1*ilvBNC*D showed that the rate-limiting sites in this chain are 1) reactions catalyzed by the AHAS and BCAT enzymes and 2) transport of valine from cells by BrnFE (Magnus et al., 2009).

Creation of valine-producing strains based on *C. glutamicum*

The information obtained so far on the biochemical, genetic, and regulatory aspects of valine biosynthesis in *C. glutamicum* suggests that the barriers to increasing valine production in this microorganism are:

- negative regulation of AHAS activity by valine, leucine, isoleucine, and 2-ketoisovalerate (retroinhibition);
- low substrate specificity of AHAS to pyruvate;
- negative regulation of *ilvBNC* operon expression by BCAA;
- consumption of pyruvate for synthesis of isoleucine, leucine, and pantothenate; and consumption of 2-ketoisovalerate for synthesis of the latter two compounds;

- expenditure of pyruvate and its precursor FEP, key metabolites of glycolytic processes, in cell energy metabolism and carboxylic acid synthesis, as well as in alanine formation;
- necessity of NADPH for the second and fourth reactions of valine biosynthesis;
- low efficiency of the BCAA BrnFE export system with respect to valine.

In the following, we will review the approaches to overcome these obstacles used in the creation of valine-producing strains based on *C. glutamicum* (information on the strains is presented in the Table).

Enhancement of valine biosynthesis reactions

Increase in AHAS activity. There are several approaches to increasing AHAS activity when creating valine-producing strains. The key one is modification of the *ilvN* gene, which eliminates retroinhibition of the enzyme. A number of mutations in the sequence of the *ilvN* gene have been found to weaken the effect of BCAA on AHAS activity. These mutations include substitutions of three amino acids, Gly20Asp, Ile21Asp, and Ile22Phe, in the *IlvN* regulatory subunit (Elišáková et al., 2005). The Ile22Phe substitution showed the best effect in this series, which was later used in other studies (Hou et al., 2012a, b). Similar effects were demonstrated for mutations leading to Ala42Val, Ala89Val, and Lys136Glu substitutions in the small subunit. The double Ala42Val-Ala89Val mutation resulted in almost complete resistance of the enzyme to inhibition by all three BCAAs (Guo et al., 2014).

Enhancement of AHAS substrate specificity with respect to pyruvate. This approach is related to the possibility of modifications of the catalytic subunit *IlvB* of AHAS that increase the affinity of the enzyme for pyruvate. Reliable data on suitable mutations are scarce. A mutation was found in the *ilvB* gene that leads to a replacement of alanine for valine at position 138 of the large AHAS subunit. This mutation has made possible a 2.5-fold increase in valine production (Liu et al., 2019). It is assumed that this substitution leads to a change in the substrate specificity of AHAS with respect to pyruvate. The molecular mechanism of action of the mutation remains unclear.

Other mutations in the *ilvB* gene of the catalytic subunit of AHAS leading to an increase in the enzyme activity toward valine production are also known (Chen et al., 2015; Guo et al., 2015). These other mutations have not yet found practical application.

The modified AHAS enzyme can be introduced into *C. glutamicum* cells in two ways: either the cells are transformed with a plasmid carrying a mutant gene (Hasegawa et al., 2012; Hou et al., 2012b; Buchholz et al., 2013) or appropriate changes are made in chromosomal DNA (Bartek et al., 2010; Hasegawa et al., 2013). Such manipulations result in a 2–3-fold increase in valine production. The use of autonomous expression plasmids also makes it possible to increase AHAS activity by introducing additional copies of the *ilvBN* genes or the entire *ilvBNC* operon into cells. The latter leads to an increase in the activity of not only AHAS, but also AHAIR.

Overcoming the negative effect of BCCA on the expression of the *ilvBNC* operon. The most rational approach to solve this problem is overexpression of the *ilvBNC* operon

using expression plasmids. At present, overexpression of the *ilvBNCDE* genes, in various combinations, is performed using constructs with strong constitutive promoters. These include, for example, promoters of superoxide dismutase (*Psod*) and elongation factor Tu (*Ptuf*) genes and a synthetic construct based on *trp* and *lac* operon promoters (*Ptac*) (Tarutina et al., 2016; Wei et al., 2018; Li et al., 2020b; Wang et al., 2020; Zhang et al., 2021). Other efficient promoters have also been described (Tarutina et al., 2016; Wei et al., 2018; Li et al., 2020b). Modifications of this type lead to an increase in valine production by about 60 % (Wei et al., 2018).

Optimization of the activity of DHAD and BCAT, which catalyze the last steps of valine biosynthesis, is provided by overexpression of the genes encoding these enzymes (*ilvD* and *ilvE*, respectively), which is usually achieved by gene amplification on plasmids (see Table). For valine production, it is especially important to increase BCAT activity because this enzyme catalyzes the rate-limiting step of biosynthesis (Magnus et al., 2009).

Minimizing the formation of byproducts

Enzymes of the valine biosynthesis pathway are involved in the formation of other metabolites such as isoleucine, leucine, and pantothenate (see Fig. 1). Consequently, activation of these enzymes and increased expression of the genes encoding them increase the yield of all the above compounds. This leads to contamination of the target product as well as a decrease in the availability of cofactors, intermediates, and the enzymes themselves for valine production. As a result, it also leads to a lower yield. Minimizing the formation of byproducts when creating strain-producers requires suppression of the relevant metabolic pathways while preserving the strains' ability to grow on poor media.

Minimization of isoleucine formation. As noted above, the synthesis of isoleucine (2-amino-3-methylpentanoic acid) is catalyzed by the same enzymes that are involved in valine biosynthesis and begins with the condensation of pyruvate and 2-ketobutyrate (see Fig. 1). The obvious way to minimize isoleucine formation is to decrease the concentration of 2-ketobutyrate in cells, the interaction of which with pyruvate determines the direction of further reactions. 2-ketobutyrate is formed from threonine by the threonine dehydratase TD, which is encoded by the *ilvA* gene (Cordes et al., 1992). The threonine dehydratase is negatively allosterically regulated by isoleucine and positively regulated by valine (Möckel et al., 1992).

The most common modification of this gene in the creation of valine-producing strains is its inactivation by deletion ($\Delta ilvA$). Most strains were obtained using this modification (see Table). It results in the appearance of the strains' ability to produce valine or a significant increase in the existing production. In this case, isoleucine auxotrophy occurs, requiring the addition of isoleucine to the cultivation medium, which complicates the production process and may increase the cost of production. In a number of studies to create valine-producing strains, instead of complete inactivation of the *ilvA* gene, a directed modification of its promoter was performed. This has resulted in a decrease in gene expression, isoleucine bradytrophism, and, as a consequence, increased production of valine (Holátko et al., 2009; Hou et al., 2012a).

Valine-producing strains engineered from *C. glutamicum*

Strain	L-Valine, g/L*	Yield, mol/mol**	References
<i>C. glutamicum</i> ATCC 13032			
$\Delta ilvA \Delta panBC$ (pJC1- <i>ilvBNCD</i>) 92 NA	10.7	–	Radmacher et al., 2002
$\Delta ilvA \Delta panBC ilvNM13$ (pECKA- <i>ilvBNC</i>)	15.2	–	Elišáková et al., 2005
$\Delta panB ilvNM13$ (P- <i>ilvAM1CG</i> P- <i>ilvDM7</i> P- <i>ilvEM6</i>)	15.9	–	Holátko et al., 2009
$\Delta aceE$ (pJC4- <i>ilvBNCE</i>)	24.6	0.60	Blombach et al., 2007
$\Delta aceE \Delta pqr$ (pJC4- <i>ilvBNCE</i>)	26.4	0.52	Blombach et al., 2008
$\Delta aceE \Delta pqr \Delta pgi$ (pJC4- <i>ilvBNCE</i>)	48.2	0.75	»
$\Delta aceE \Delta pqr \Delta pgi \Delta pyc$ (pJC4- <i>ilvBNCE</i>)	28.1	0.86	»
$\Delta aceE \Delta pqr$ (pJC4- <i>ilvBNCE</i>)	24.6	0.23	Blombach et al., 2009
$\Delta aceE \Delta pqr \Delta sugR$ (pJC4- <i>ilvBNCE</i>)	35.2	0.20	»
$\Delta aceE \Delta pqr$ (pJC4- <i>ilvBNCE</i>) (pBB1- <i>pntAB</i>)	14.6	0.92	Bartek et al., 2011
(P- <i>ilvAM1CG</i>) $\Delta avtA$ pDXW-8- <i>ilvEBN(r)C</i>	31.2	0.17	Hou et al., 2012a
$\Delta ilvA \Delta panB ilvNM13$ (pECKA- <i>ilvBNC</i>)	12.5	–	Denina et al., 2010
$\Delta ilvA \Delta panB \Delta rel ilvNM13$ (pECKA- <i>ilvBNC</i>)	11.5	–	»
$\Delta ilvA \Delta panBC \Delta avtA$ (pJC4- <i>ilvBNCE</i>)	8.8	–	Marienhagen, Eggeling, 2008
$aceEA16 \Delta pqr \Delta ppc$ (pJC4- <i>ilvBNCE</i>)	86.5	0.36	Buchholz et al., 2013
$\Delta ppc \Delta pyc icd^{Ala94Asp}$ (pJC4- <i>ilvBNCE</i>)	8.0	0.20	Schwentner et al., 2018
$\Delta ppc \Delta pyc icd^{Gly407Ser}$ (pJC4- <i>ilvBNCE</i>)	8.9	0.22	»
$\Delta ponA \Delta ilvA P ilvB^{G183A}$	15.6	–	Ryabchenko et al., 2021
<i>C. glutamicum</i> ATCC 13869			
$\Delta aceE \Delta alaT \Delta ilvA$ (pJYW4- <i>ilvBNC1-lrp1-brnFE</i>)	51.2	0.47	Chen et al., 2015
$\Delta ponA \Delta ilvA P ilvB^{G183A}$	25.1	–	Ryabchenko et al., 2021
<i>C. glutamicum</i> R			
$\Delta ldhA$ (pCRB-BN ^{GE} C TM) (pCRB-DLD)	172.2***	0.63	Hasegawa et al., 2012
$\Delta ldhA \Delta ppc \Delta pta \Delta ackA \Delta ctgA \Delta avtA ilvN^{GE}C^{TM} + gapA+pyk+pfkA+pgi+tpi$ (pCRB-BN ^{GE} C TM) (pCRB-DLD)	149.9***	0.88	Hasegawa et al., 2013
<i>B. flavum</i> ATCC14067 (<i>C. glutamicum</i> ssp. <i>flavum</i>)			
pDXW-8- <i>ilvEBN(r)</i>	38.1	0.24	Hou et al., 2012b
$\Delta alr \Delta aceE \Delta ilvA \Delta leuA$ (pJYW4)	9.5	–	Liu et al., 2019
$\Delta alr \Delta aceE \Delta ilvA \Delta leuA$ (pJYW4- <i>ilvB</i> ^{138Val404Ala-ilvN})	14.5	–	»
$\Delta alr \Delta aceE \Delta ilvA \Delta leuA$ (pJYW4- <i>ilvB</i> ^{138Val404Ala-ilvNCE})	25.9	0.49	»
<i>B. flavum</i> JV16			
<i>avtA::Cm</i> (pDXW8- <i>ilvEBN(r)C</i>)	34.4	0.22	Hou et al., 2012a
<i>C. glutamicum</i>			
$\Delta ppc \Delta aceE \Delta alaT \Delta pqr$	3.2	–	Han et al., 2020

* Concentration of valine in the culture liquid.

** Yield of target product (valine) from substrate (glucose).

*** Data were obtained using concentrated cell suspension.

Another target for modifications aimed at reducing isoleucine biosynthesis is AHAS. A variant modification of the enzyme's catalytic subunit that increases its specificity toward pyruvate and redirects cellular resources toward valine production (Liu et al., 2019) is described above.

Minimization of leucine and pantothenate formation. 2-ketoisovalerate is a precursor not only to valine but also to leucine and pantothenate (see Fig. 1). The synthesis of leucine (2-amino-4-methylpentanoic acid) from 2-ketoisovalerate is controlled by the *leuA*, *leuB*, and *leuCD* genes localized in different regions of the chromosome. It is known that *leuB* and *leuCD* are subject to the control of the LtbR transcriptional repressor, while *leuA* regulation seems to involve the mechanism of attenuation of transcription (Wang et al., 2019a). A modification to preserve 2-oxoisovalerate for valine biosynthesis at the expense of decreased leucine biosynthesis was carried out by J. Holátko and colleagues (2009) by reducing the expression of the *leuA* gene by replacing the native promoter with a weaker one. The result was a 50–70 % increase in valine production, which is comparable to the effect observed when the expression of the *ilvA* gene is weakened.

The synthesis of pantothenate (amide of β -alanine and pantoic acid) from 2-ketoisovalerate is controlled by the *panB* and *panC* genes, which form one operon (Sahm, Eggeling, 1999), and the *panD* gene which is located separately in the genome (Dusch et al., 1999). It was noted that the carbon flux going to valine biosynthesis is 10 times higher than the flux going to pantothenate biosynthesis, even in the strain with enhanced expression of *panBC* (Chassagnole et al., 2002). However, inactivation of the *panB* gene or the entire *panBC* operon has a favorable effect on valine production, even though it leads to pantothenate auxotrophy in strains. This inactivation allows valine production to appear in wild-type strains and to increase valine production in valine-producing strains by more than 30 % or even 50 % (Radmacher et al., 2002; Holátko et al., 2009).

Increasing availability of precursors and cofactors

Increasing availability of pyruvate. Pyruvate, the centerpiece of carbon and energy metabolism in all organisms, is a precursor not only to BCAA and pantothenate, but also to many other compounds, including components of the tricarboxylic acid cycle (TCA cycle) as well as lactate and alanine (see Fig. 1). Efficient production of valine requires maintaining a pool of pyruvate in the cells and, therefore, enhancing pyruvate formation reactions and/or reducing its “off-target” consumption. Pyruvate, which synthesizes 2 mol of reduced nicotinamide dinucleotide (NADH), is a product of glycolysis (Wieschalka et al., 2012). However, glycolytic enzyme activity is generally not increased in the development of valine producers, except for the microaerobic process (see below). The main approach is to reduce the outflow of pyruvate, and its precursor PEP, into other pathways.

One of the main pathways of pyruvate outflow is the TCA cycle. This process becomes less active in the late stages of growth, which could be used to create favorable conditions for valine production. Indeed, a decrease in the growth rate of *C. glutamicum* is accompanied by an increase in pyruvate levels in cells and an increase in valine (Ruklisha et al., 2007). In valine-producing strains that are auxotrophic

for isoleucine and pantothenate, growth of cultures can be controlled by changing the amount of supplementation with these substances. Growth restriction also leads to increased productivity (Bartek et al., 2008).

Involvement of pyruvate and PEP in the TCA cycle occurs both through conversion of both compounds to oxaloacetate (OA) and of pyruvate to acetyl-CoA directly or through acetate and acetyl-phosphate (see Fig. 1). As a rule, increasing the production of valine as well as pyruvate itself is attempted through reducing the activity of the pyruvate dehydrogenase complex (PDHC), which catalyzes the oxidative decarboxylation of pyruvate to acetyl-CoA. In *C. glutamicum*, this complex consists of three subunits, E1, E2, and E3, encoded by the *aceE*, *aceF*, and *lpd* genes, respectively (Eikmanns, Blombach, 2014). Inactivation of the *aceE* gene by deletion ($\Delta aceE$) is one of the most frequent steps in creating a valine producer (see Table). The resulting strains require the addition of acetate in minimal medium, but the level of valine production increases manifold. Metabolomic analysis showed that inactivation of *aceE* in wild-type *C. glutamicum* leads to a 13-fold increase in the pyruvate pool in cells (Blombach et al., 2007).

A characteristic feature of *C. glutamicum* strains devoid of PDHC is the production of valine in the absence of cell growth. Increased glucose utilization rate was achieved by adding maltose instead of glucose, using ethanol instead of acetate, or inactivating the transcriptional regulator SugR (Blombach et al., 2009; Krause et al., 2010b). SugR in *C. glutamicum* is responsible for acetate-mediated repression of the *ptsG*, *ptsI*, and *ptsH* genes encoding the enzymes of the phosphotransferase system (PTS). PTS ensures the conjugated processes of sugars transport into the cell and their phosphorylation (Engels, Wendisch, 2007). However, because of PDHC deficiency, all strains still needed acetate or ethanol, which is then also converted to acetate as an additional carbon source.

To overcome this need, the native *aceE* gene promoter was replaced with mutant variants from a previously established promoter library based on the *dapA* gene promoter (Vasicová et al., 1999). This allowed to obtain a series of *C. glutamicum* strains with gradually decreased PDHC activity as well as gradually decreased growth rate on medium containing glucose as the only carbon source. Transformed with the *pJC4-ilvBNCE* plasmid, these strains produced valine and did not require acetate as an additional carbon source (Buchholz et al., 2013). A growth-dependent promoter of the aldehyde dehydrogenase gene from *C. glutamicum* CP (P_{CP_2836}) has been used for the same purposes. This has led to a threefold decrease in *aceE* transcription levels compared to the native promoter, as well as has had positive effects on both cell growth and valine production (Ma et al., 2018b).

It is also possible to reduce pyruvate consumption in the TCA cycle by decreasing the activity of the cycle itself. For example, suppression of the gene of the transcription factor RamA responsible for the TCA cycle activation has been shown to contribute to efficient pyruvate production (Kataoka et al., 2019).

The conversion of pyruvate to acetate is catalyzed by pyruvate:quinoxido-reductase (product of the *pqo* gene), the inactivation of which (Δpqo) leads to increased valine production (see Table), but also to impaired growth characteristics of

strains. The combination of this modification with inactivation of PEP carboxylase (product of the *ppc* gene), which catalyzes formation of OA from PEP, resulted in a slight increase in valine production, however, the yield increased by 14 % (Buchholz et al., 2013). It was noted that the valine-producing strain with inactivated *aceE* and *pqo* genes grew better and produced more valine on maltose-enriched medium (Krause et al., 2010b).

Another pathway for the outflow of pyruvate is the formation of OA from it under the action of pyruvate carboxylase (product of the *pyc* gene). Inactivation of *pyc* in the creation of a valine-producing strain leads to an increase in yield to 0.86 mol of valine per 1 mol of glucose (Blombach et al., 2008). When developing a leucine-producing strain, it was found that, in order to minimize pyruvate outflow, inactivation of pyruvate carboxylase is more beneficial than inactivation of PEP carboxylase (Wang et al., 2020).

Two other pathways of pyruvate consumption in *C. glutamicum* cells are the processes of lactate and alanine biosynthesis (see Fig. 1). Lactate formation catalyzed by lactate dehydrogenase (a product of the *ldhA* gene) becomes important in terms of valine production under oxygen deficiency conditions (Hasegawa et al., 2012) and will be discussed further.

Minimization of alanine synthesis is required under all conditions because this process leads not only to untargeted consumption of pyruvate but also to loss of NADPH in the amino group transfer reaction and to unwanted impurities in the final product.

Alanine formation in *Corynebacteria* is catalyzed by the transaminases AlaT and AvtA, which use glutamate and valine as amino group donors, respectively (Marienhagen et al., 2005; Marienhagen, Eggeling, 2008). It was noted above that AvtA is one of the transaminases involved in valine biosynthesis, but its role, compared with BCAT, is minor.

The question of the participation of these transaminases in alanine biosynthesis in *C. glutamicum* remains open due to the inconsistency of existing data. On the one hand, inactivation of *alaT* and *avtA* in the valine-producing strain has been shown to reduce alanine formation by about 80 and 20 %, respectively (Marienhagen, Eggeling, 2008). A significant decrease in alanine synthesis (to less than 0.2 g/L) is observed as a result of the inactivation of both genes (Hou et al., 2012a). These data suggest that the AlaT aminotransferase is the major one, but both enzymes are involved in alanine synthesis. On the other hand, in the proline producer, inactivation of *alaT* has no effect on alanine levels, whereas inactivation of *avtA* reduces this level by 48 % (Zhang et al., 2020). Moreover, analysis of the transcriptome of the industrial valine producer line VWB-1 showed that its low level of L-alanine synthesis is not associated with the *alaT* gene, the transcriptional level of which in this strain is 5.1-fold higher than that in the wild-type strain. It is assumed that a lower level of L-alanine synthesis is due to the lower expression of the gene *alr* encoding alanine racemase that converts L-alanine to D-alanine (Zhang H. et al., 2018). Thus, it is also impossible to give an unequivocal answer to the question of whether inactivation of one or the other of these two transaminases is more advantageous in terms of valine production.

Increasing availability of NADPH. In *C. glutamicum*, the main supplier of NADPH is PPP, in which the reduction of

NADP⁺ to NADPH is provided by glucose-6-phosphate dehydrogenase (a heteromultimeric complex wherein one of the subunits is encoded by the *zwf* gene) and 6-phosphogluconate dehydrogenase (a product of the *gnd* gene). The activity of both enzymes is negatively regulated by ATP, NADPH, and other metabolites (Moritz et al., 2000). NADPH-dependent decarboxylating malate dehydrogenase (malic enzyme) and isocitrate dehydrogenase play a minor role in the synthesis of NADPH from NADP⁺ (Bartek et al., 2010; Siedler et al., 2013). The source of NADP⁺ and, hence, the source of NADPH in *C. glutamicum* can also be NAD⁺, which is phosphorylated by NAD kinase (product of the *ppnK* gene) to form NADP⁺. This enzyme has been characterized as a polyphosphate-ATP-dependent NAD kinase that uses ATP to phosphorylate NAD⁺ (Shi et al., 2013).

Theoretical analysis showed that the level of substrate conversion to valine (the yield) significantly depends on the reactions used for NADPH regeneration. The maximum yield, equal to 1 mol of valine per 1 mol of glucose, is obtained without the expenditure of carbon for growth and synthesis of NADPH. If NADPH is provided by isocitrate dehydrogenase activity, the yield is 0.5 mol of valine per 1 mol of glucose. Directing the entire carbon flux into the NADPH-generating PPP results in a much higher yield of 0.86. In this analysis, the main target for the redirection of carbon flux from the TCA cycle to valine biosynthesis appeared to be PDHC. A scenario in which carbon is not consumed for NADPH synthesis at all can be realized by the combined activity of pyruvate carboxylase (or PEP carboxylase), malate dehydrogenase, and malic enzyme, theoretically capable of transferring hydrogen from NADH to NADP⁺ (Bartek et al., 2010). Such a pathway, designated a transhydrogenase-like shunt, is involved in NADPH formation for anaerobic isobutanol production in *C. glutamicum* (Blombach, Eikmanns, 2011). Thus, enhancement of PPP and NAD kinase activity are the most obvious ways to increase the NADPH pool in the cell.

From the point of view of the efficiency of the valine biosynthesis process, it is advantageous to combine the enhancement of PPP with some weakening of glycolysis. Indeed, inactivation of the glucose-6-phosphatase gene *pgi* (this inactivation directs carbon flux from glycolysis to PPP) resulted in more efficient valine production in the *C. glutamicum* strain $\Delta aceE \Delta pqo \Delta pgi$ (*pilvBNCE*), producing 48.0 g/L with a yield of 0.75 mol of valine per 1 mol of glucose (Blombach et al., 2008). Further analysis of this strain showed that inactivation of *pgi* results in increased intracellular NADPH concentrations and decreased byproduct formation (Bartek et al., 2010). Monitoring cellular NADPH content using NADPH-dependent fluorescence also showed that the *C. glutamicum* strain carrying Δpgi does accumulate NADPH (Goldbeck et al., 2018).

The growth deterioration observed in Δpgi -strains on medium with glucose has been attributed to a decrease in PTS activity and suggested to be overcome by overexpression of the gene *ptsG*, which encodes a glucose-specific component of this system (Lindner et al., 2013). For *pgi*-mutants, enhancement of the alternative glucose transport system by inositol permeases IolT1, IolT2, and the glucokinase PpgK, which was used to produce lysine producer, is also effective (Xu J.Z. et al., 2019).

Another approach to increase the NADPH pool is related to the possibility of changing the specificity of glycolytic enzymes from NAD^+ to NADP^+ . It has been implemented to improve lysine production. Point mutations in the glyceraldehyde-3-phosphate dehydrogenase *gapA* gene that changed enzyme specificity resulted in a 35–60 % increase in lysine production (Bommareddy et al., 2014; Xu et al., 2014).

It was noted above that enzymes that synthesize NADPH are susceptible to negative regulation by various metabolites. Therefore, one approach to PPP activation is to introduce into the corresponding genes mutations that increase enzyme activity. Such an approach has been implemented for the *zwf* and *gnd* genes in works on methionine, proline, and riboflavin producers. It has indeed led to an increase in the NADPH pool and production levels in cells (Wang et al., 2011; Li et al., 2016; Zhang et al., 2020).

As for NAD-kinase, the studies published to date on enhancing its activity target isoleucine production. These studies indicate that modifications that increase the enzyme activity (point mutations in the *ppnK* gene, overexpression of the *ppnK* gene) lead to increased intracellular concentration of NADP^+ and NADPH and contribute to enhanced biosynthesis of the target product (Yin et al., 2014; Zhang et al., 2020).

Another attractive possibility for increasing NADPH availability for valine biosynthesis is heterologous expression of transhydrogenase genes, such as *pntAB* from *E. coli*, that catalyze NADP^+ reduction involving NADH. This possibility was previously used to improve lysine production with *C. glutamicum* (Kabus et al., 2007). A significant increase in intracellular NADPH concentration was observed when *pntAB* expression was combined with overexpression of the *ppnK* gene (Zhan et al., 2019). Introduction of PntAB from *E. coli* into the valine-producing strain *C. glutamicum* $\Delta\text{aceE} \Delta\text{pqr}$ (pJC4*ilvBNCE*) resulted in a significant decrease in carbon flux in PPP and, consequently, an increase in yield to 0.92. This is the highest yield (Bartek et al., 2011), which is only 8 % below the theoretical maximum of 1 mol of valine per 1 mol of glucose (Bartek et al., 2010).

Engineering the microaerobic process of valine production

Under oxygen deprivation, *C. glutamicum* cultures show very poor growth capacity but metabolize sugars to organic acids (Michel et al., 2015; Lange et al., 2018). When byproduct synthesis is suppressed, producer strains adapted to such conditions show higher productivity than strains requiring aeration (Okino et al., 2008; Jojima et al., 2010, 2015; Yamamoto et al., 2013). Valine biosynthesis under normal conditions is an aerobic process because it is carried out by growing cultures actively generating NADPH. For efficient production of valine under oxygen deprivation, strains require a complex modification involving both valine biosynthesis enzymes and glycolysis enzymes. Such a modification was performed by S. Hasegawa and colleagues (2012, 2013).

The *C. glutamicum* R strain with inactivated lactate dehydrogenase (ΔldhA) and overexpression of the *ilvBNCE* genes encoding the enzymes of valine biosynthesis was used as the basis for creating strains producing valine under microaerobic conditions. This strain is incapable of producing valine under oxygen deficiency because it has an imbalance of cofactors:

2 mol of NADPH are consumed while 2 mol of NADH are synthesized to produce 1 mole of valine.

The appearance of valine production was achieved by using two approaches. The first approach was to change the specificity of AHAS from NADPH to NADH by site-directed mutagenesis of the *ilvC* gene (constructing the *ilvCTM* gene). The second approach was to replace the NADPH-dependent transaminase BCAT with NAD-dependent leucine dehydrogenase (LeuDH) from *Lysinibacillus sphaericus* (Hasegawa et al., 2012). The additional introduction of the *ilvN* gene encoding a mutant AHAS regulatory subunit (*ilvN^{GE}*) resistant to BCAA inhibition has allowed to produce a *C. glutamicum* strain (pCRB-*BN^{GE}CTM*)(pDLD)/ ΔLDH) that produced 172.2 g/L of valine for 24 h under microaerobic conditions with periodic fermentation, which was more than 20-fold higher than baseline. The yield was 0.63 mol of valine per mol of glucose (Hasegawa et al., 2012).

However, in addition to valine, the cells of this strain accumulated significant amounts of alanine, acetate, and succinate as byproducts. To eliminate their formation and increase the valine yield, additional modifications were introduced into the strain (Hasegawa et al., 2013). Succinate formation via PEP and OA was suppressed by inactivation of the *ppc* gene, but this resulted in reduced valine synthesis and glucose uptake, as the intracellular NADH/ NAD^+ ratio increased markedly. To restore the ratio to a level favorable for valine production, three genes involved in acetate synthesis (*pta*, *ackA*, *ctfA*) were inactivated and the expression of five genes (*gapA*, *pyk*, *pfkA*, *pgi*, *tpi*) encoding glycolysis enzymes was increased. As a result, valine production increased 9-fold and glucose uptake, 7.6-fold. Since valine biosynthesis became an NADH-dependent process, increasing the activity of glycolytic enzymes turned out to be beneficial in terms of accumulating both pyruvate and reducing equivalents.

Decrease in alanine formation was achieved by inactivation of the *avtA* gene. In addition, the *ilvN^{GE}* and *ilvCTM* genes, which were previously expressed on the plasmid, were placed in the chromosome. The valine productivity of the new strain was 149.9 g/L in 24 h of cultivation. The yield reached 0.88 mol of valine per mol of glucose, which was significantly higher than that obtained in the first step (Hasegawa et al., 2013).

It should be noted that in both works, valine synthesis under microaerobic conditions was studied using non-growing cells preconcentrated by centrifugation by a factor of 2 to 3. In this case, the measured valine concentration reached very high values, but the productivity per cell was comparable with that demonstrated in other studies.

Replacement of enzyme specificity from NADPH to NADH to adapt the amino acid production process to microaerobic conditions has also been done in the development of *E. coli*-based valine producer (Savrasova, Stoyanova, 2019) and *C. glutamicum*-based leucine and L-ornithine producers (Jiang et al., 2013; Wang et al., 2019b). In all cases, this resulted in an increased yield of the target product.

The engineering of valine transport

Microorganisms have multiple transport systems that ensure the uptake of desired environmental components by cells and release of metabolites, the excess of which can be toxic

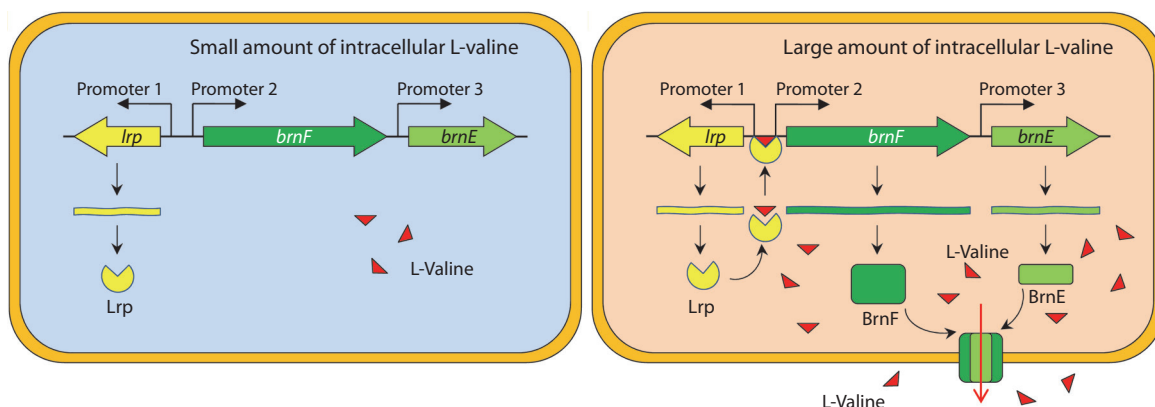


Fig. 3. Organization of the *brnFE* operon in *C. glutamicum* and regulation of its expression (from the review (Wang et al., 2018)).

(Pérez-García, Wendisch, 2018). The activity of such systems depends on the concentration of the transported substances, so it has long been thought that producing strains' own regulatory mechanisms are sufficient for excreting the target products effectively (Jones et al., 2015). Transport engineering is complicated by the complexity of its quantification and the fact that specific transporters are not known for each biotechnologically relevant substance. In recent years, however, there have been a growing number of studies showing the effect of directional changes in export and import of the target product on strain productivity (Eggeling, 2016). Valine transporters in *Corynebacteria* have been detected and characterized, and thus are promising targets for modifications in the creation of producing strains.

Valine import. The uptake of valine, leucine, and isoleucine in *Corynebacteria* occurs through a secondary Na^+ -dependent symport carried out by the only known importer, BrnQ (Ebbighausen et al., 1989). BrnQ exhibits the highest affinity for isoleucine. For valine and leucine, the affinity is 1.7 times lower (Ebbighausen et al., 1989; Tauch et al., 1998). Data on the regulation of BrnQ and the corresponding gene in *corynebacteria* are extremely scarce. It is known that BrnQ is activated when the intracellular concentration of BCAA is increased (Boles et al., 1993) and that inactivation of the *brnQ* gene increases isoleucine export from *C. glutamicum* cells and its production (Xie et al., 2012). It has been noted that a similar modification favors growth and productivity of the isoleucine-producing strain WM001 in the early stages of fermentation (Zhang et al., 2020). The importance of the importer for valine production is confirmed by transcriptome analysis of the industrial producer VWB-1, which showed that the transcription level of the *brnQ* gene in this strain is lower than that of the wild-type strain (Zhang H. et al., 2018).

Valine export. The BrnFE transport system is responsible for BCAA export from *C. glutamicum* cells (Eggeling, Sahm, 2003). Amino acids are exported through a secondary H^+ -dependent process controlled by membrane potential (Hermann, Kramer, 1996). The *brnFE* transport system is the only known exporter of valine, leucine, and isoleucine in *C. glutamicum*. It also transports methionine and homoserine, a precursor of methionine, isoleucine, and threonine (Kennerknecht et al., 2002; Trotschel et al., 2005; Yin et al., 2013; Qin et al., 2015; Li et al., 2020a). The *brnF* and *brnE* genes encoding,

respectively, the large and small subunits of the transport system, are organized into a single operon controlled by the transcriptional regulator Lrp (leucine responsive protein) (Kennerknecht et al., 2002; Lange et al., 2012). Homologues of Lrp, first discovered and characterized in *E. coli*, are present in the genomes of various prokaryotes and regulate genes involved in amino acid metabolism (Brinkman et al., 2003). In *C. glutamicum*, the *lrp* gene is located divergently upstream of the *brnFE* operon. By binding to BCAA or methionine, Lrp becomes active and, in turn, activates the *brnFE* promoter (Kennerknecht et al., 2002; Lange et al., 2012) (Fig. 3). The effect of cellular amino acid concentration on Lrp activity decreases in the series leucine > methionine > isoleucine > valine (Lange et al., 2012).

A study of industrial leucine and valine producers confirms that a high level of amino acid production either correlates with a high level of operon *brnFE* expression (Vogt et al., 2014; Zhang H. et al., 2018) or is associated with an increased *lrp* and *brnFE* gene copy number (Ma et al., 2018a).

Analysis of the effect of operon *brnFE* on valine production showed that its overexpression does not affect the growth of *C. glutamicum* cells and increases valine production by about 2–3 times (Chen et al., 2015). Overexpression of *brnFE* has a similar effect on the production of isoleucine, methionine, and homoserine (Qin et al., 2015; Li et al., 2020a; Zhang et al., 2021). The maximum effect on isoleucine production was obtained when *lrp* and *brnFE* expression were simultaneously enhanced (Yin et al., 2013).

However, it was found that, unlike *brnFE*, overexpression of the *lrp* gene suppresses cell growth (Chen et al., 2015), although it also significantly increases valine production. The negative effect was counterbalanced by the use of a weakened mutant form of this *lrp*₁ gene found in the VWB-1 strain. Overexpression of *lrp*₁ in the wild-type *C. glutamicum* strain resulted in a 16-fold increase in valine production, from 1.9 to 30.2 mmol/L per 96 h of cultivation. The combination of *lrp*₁ and *brnFE* overexpression enhanced the effect. Isoleucine production was not significantly affected by such manipulations, from which the authors concluded that isoleucine is a less suitable substrate for *brnFE* than valine (Chen et al., 2015). Simultaneous amplification of the expression of both genes, *lrp* and *brnFE*, combined with overexpression of the *ilvBNC* genes and inactivation of *aceE*, *alaT*, and *ilvA*, resulted in a

strain that produced 437 mM (51 g/L) valine when fermented with feeding (Chen et al., 2015).

Thus, modifications of BCAA transport systems aimed at reducing the influx of amino acids into the cell and increasing their secretion from the cell have a positive effect on the production of the amino acids (Xie et al., 2012).

Conclusion

In recent years, interest in the use of valine as a feed additive has increased significantly. In the Russian Federation alone, the consumption of valine has increased almost 10-fold over the past five years, reaching 5,000 tons per year. Modern industrial production of valine is based on microbiological synthesis using renewable plant raw materials and producing strains with a modified genetic program. The efficiency of amino acid production largely depends on the productivity of the producer strains, which are a key element of the entire process chain. Although significant progress has been made in the creation of producing strains (see Table), the creation of new strains with unique characteristics is still relevant.

It is worth noting that the recently developed processes with reduced aeration have a higher potential compared to the traditional aerobic processes of valine production. However, it should be noted that such processes are biphasic: in the first phase, biomass is produced aerobically, while in the second phase, valine biosynthesis occurs under microaerobic conditions. Currently, the two-phase processes show low efficiency, and more research in this area is required.

Nowadays, the main approach to creating valine-producing strains, which has replaced random mutagenesis, is rational metabolic engineering aimed at enhancing the valine biosynthesis process and minimizing the formation of byproducts. In recent years, this approach has been actively enriched by the application of systems engineering and synthetic biology methods. The combined analysis of “omics” data expands our knowledge of the metabolic and regulatory processes of *C. glutamicum* and allows us to develop new strategies for creating producers of valine and other amino acids. The recent emergence of rapid genome editing systems that speed up the process of obtaining new strains should help to implement these strategies.

Further progress in the creation of producer strains will involve a shift from studying the properties of a cell population to studying the properties of individual cells (Harst et al., 2017; Hemmerich et al., 2018; Pérez-García et al., 2018), as well as extensive application of computer modeling (Koduru et al., 2018) and using new knowledge about gene expression regulation (Dostálová et al., 2017; Shi et al., 2018; Zhang S. et al., 2018; Xu N. et al., 2019).

The approaches perfected in the creation and improvement of valine producers can be used to create producers of other BCAA and pantothenate, the substances that also have a significant market potential.

References

- Baritugo K.A., Kim H.T., David Y., Choi J.I., Hong S.H., Jeong K.J., Choi J.H., Joo J.C., Park S.J. Metabolic engineering of *Corynebacterium glutamicum* for fermentative production of chemicals in biorefinery. *Appl. Microbiol. Biotechnol.* 2018;102(9):3915-3937. DOI 10.1007/s00253-018-8896-6.

- Bartek T., Blombach B., Lang S., Eikmanns B.J., Wiechert W., Oldiges M., Noh K., Noack S. Comparative C-13 metabolic flux analysis of pyruvate dehydrogenase complex-deficient, L-valine-producing *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 2011; 77(18):6644-6652. DOI 10.1128/aem.00575-11.
- Bartek T., Blombach B., Zonnchen E., Makus P., Lang S., Eikmanns B.J., Oldiges M. Importance of NADPH supply for improved L-valine formation in *Corynebacterium glutamicum*. *Biotechnol. Prog.* 2010;26(2):361-371. DOI 10.1002/btpr.345.
- Bartek T., Makus P., Klein B., Lang S., Oldiges M. Influence of L-isoleucine and pantothenate auxotrophy for L-valine formation in *Corynebacterium glutamicum* revisited by metabolome analyses. *Bioprocess Biosyst. Eng.* 2008;31(3):217-225. DOI 10.1007/s00449-008-0202-z.
- Blombach B., Arndt A., Auchter M., Eikmanns B.J. L-valine production during growth of pyruvate dehydrogenase complex deficient *Corynebacterium glutamicum* in the presence of ethanol or by inactivation of the transcriptional regulator SugR. *Appl. Environ. Microbiol.* 2009;75(4):1197-1200. DOI 10.1128/aem.02351-08.
- Blombach B., Eikmanns B.J. Current knowledge on isobutanol production with *Escherichia coli*, *Bacillus subtilis* and *Corynebacterium glutamicum*. *Bioeng. Bugs.* 2011;2(6):346-350. DOI 10.4161/bbug.2.6.17845.
- Blombach B., Schreiner M.E., Bartek T., Oldiges M., Eikmanns B.J. *Corynebacterium glutamicum* tailored for high-yield L-valine production. *Appl. Microbiol. Biotechnol.* 2008;79(3):471-479. DOI 10.1007/s00253-008-1444-z.
- Blombach B., Schreiner M.E., Holátko J., Bartek T., Oldiges M., Eikmanns B.J. (L)-valine production with pyruvate dehydrogenase complex-deficient *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 2007;73(7):2079-2084. DOI 10.1128/aem.02826-06.
- Boles E., Ebbighausen H., Eikmanns B., Krämer R. Unusual regulation of the uptake system for branched-chain amino acids in *Corynebacterium glutamicum*. *Arch. Microbiol.* 1993;159:147-152. DOI 10.1007/BF00250275.
- Bommareddy R.R., Chen Z., Rappert S., Zeng A.P. A *de novo* NADPH generation pathway for improving lysine production of *Corynebacterium glutamicum* by rational design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase. *Metab. Eng.* 2014;25: 30-37. DOI 10.1016/j.ymben.2014.06.005.
- Brinkman A.B., Ettema T.J., de Vos W.M., van der Oost J. The Lrp family of transcriptional regulators. *Mol. Microbiol.* 2003;48(2): 287-294. DOI 10.1046/j.1365-2958.2003.03442.x.
- Buchholz J., Schwentner A., Brunnenkan B., Gabris C., Grimm S., Gerstmeir R., Takors R., Eikmanns B.J., Blombach B. Platform engineering of *Corynebacterium glutamicum* with reduced pyruvate dehydrogenase complex activity for improved production of L-lysine, L-valine, and 2-ketoisovalerate. *Appl. Environ. Microbiol.* 2013;79(18):5566-5575. DOI 10.1128/AEM.01741-13.
- Burkovski A. I do it my way: regulation of ammonium uptake and ammonium assimilation in *Corynebacterium glutamicum*. *Arch. Microbiol.* 2003;179(2):83-88. DOI 10.1007/s00203-002-0505-4.
- Chassagnole C., Létisse F., Diano A., Lindley N.D. Carbon flux analysis in a pantothenate overproducing *Corynebacterium glutamicum* strain. *Mol. Biol. Rep.* 2002;29(1-2):129-134. DOI 10.1023/a:1020353124066.
- Che L., Xu M., Gao K., Wang L., Yang X., Wen X., Xiao H., Li M., Jiang Z. Mammary tissue proteomics in a pig model indicates that dietary valine supplementation increases milk fat content via increased *de novo* synthesis of fatty acid. *Food Sci. Nutr.* 2021;9(11): 6213-6223. DOI 10.1002/fsn3.2574.
- Chen C., Li Y., Hu J., Dong X., Wang X. Metabolic engineering of *Corynebacterium glutamicum* ATCC13869 for L-valine production. *Metab. Eng.* 2015;29:66-75. DOI 10.1016/j.ymben.2015.03.004.
- Chen X.H., Liu S.R., Peng B., Li D., Cheng Z.X., Zhu J.X., Zhang S., Peng Y.M., Li H., Zhang T.T., Peng X.X. Exogenous L-valine promotes phagocytosis to kill multidrug-resistant bacterial pathogens. *Front. Immunol.* 2017;8:207. DOI 10.3389/fimmu.2017.00207.

- Cordes C., Möckel B., Eggeling L., Sahm H. Cloning, organization and functional analysis of *ilvA*, *ilvB* and *ilvC* genes from *Corynebacterium glutamicum*. *Gene*. 1992;112(1):113-116. DOI 10.1016/0378-1119(92)90311-c.
- Denina I., Paegle L., Prouza M., Holátko J., Pátek M., Nesvera J., Ruklisha M. Factors enhancing L-valine production by the growth-limited L-isoleucine auxotrophic strain *Corynebacterium glutamicum* DeltailvA DeltapanB ilvNM13 (pECKAilvBNC). *J. Ind. Microbiol. Biotechnol.* 2010;37(7):689-699. DOI 10.1007/s10295-010-0712-y.
- D'Este M., Alvarado-Morales M., Angelidaki I. Amino acids production focusing on fermentation technologies – A review. *Biotechnol. Adv.* 2017;36(1):14-25. DOI 10.1016/j.biotechadv.2017.09.001.
- Dimou A., Tsimihodimos V., Bairaktari E. The critical role of the branched chain amino acids (BCAAs) catabolism-regulating enzymes, branched-chain aminotransferase (BCAT) and branched-chain α -keto acid dehydrogenase (BCKD), in human pathophysiology. *Int. J. Mol. Sci.* 2022;23(7):4022. DOI 10.3390/ijms23074022.
- Dostálová H., Holátko J., Busche T., Rucká L., Rapoport A., Halada P., Nešvera J., Kalinowski J., Pátek M. Assignment of sigma factors of RNA polymerase to promoters in *Corynebacterium glutamicum*. *AMB Express*. 2017;7(1):133. DOI 10.1186/s13568-017-0436-8.
- Dusch N., Pühler A., Kalinowski J. Expression of the *Corynebacterium glutamicum* *panD* gene encoding L-aspartate- α -decarboxylase leads to pantothenate overproduction in *Escherichia coli*. *Appl. Environ. Microbiol.* 1999;65(4):1530-1539. DOI 10.1128/AEM.65.4.1530-1539.1999.
- Ebbighausen H., Weil B., Krämer R. Transport of branched-chain amino acids in *Corynebacterium glutamicum*. *Arch. Microbiol.* 1989;151(3):238-244. DOI 10.1007/BF00413136.
- Eggeling L., Cordes C., Eggeling L., Sahm H. Regulation of acetohydroxy acid synthase in *Corynebacterium glutamicum* during fermentation of alpha-ketobutyrate to L-isoleucine. *Appl. Microbiol. Biotechnol.* 1987;25(4):346-351. DOI 10.1007/BF00252545.
- Eggeling L. Exporters for production of amino acids and other small molecules. *Adv. Biochem. Eng. Biotechnol.* 2016;159:199-225. DOI 10.1007/10_2016_32.
- Eggeling L., Sahm H. New ubiquitous translocators: amino acid export by *Corynebacterium glutamicum* and *Escherichia coli*. *Arch. Microbiol.* 2003;180(3):155-160. DOI 10.1007/s00203-003-0581-0.
- Eikmanns B., Blombach B. The pyruvate dehydrogenase complex of *Corynebacterium glutamicum*: an attractive target for metabolic engineering. *J. Biotechnol.* 2014;192(Pt. B):339-345. DOI 10.1016/j.jbiotec.2013.12.019.
- Elišáková V., Pátek M., Holátko J., Nesvera J.N., Leyval D., Goergen J.L., Delaunay S. Feedback-resistant acetohydroxy acid synthase increases valine production in *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 2005;71(1):207-213. DOI 10.1128/aem.71.1.207-213.2005.
- Engels V., Wendisch V.F. The DeoR-type regulator SugR represses expression of *ptsG* in *Corynebacterium glutamicum*. *J. Bacteriol.* 2007;189(8):2955-2966. DOI 10.1128/JB.01596-06.
- Goldbeck O., Eck A.W., Seibold G.M. Real time monitoring of NADPH concentrations in *Corynebacterium glutamicum* and *Escherichia coli* via the genetically encoded sensor mBFP. *Front. Microbiol.* 2018;9:2564. DOI 10.3389/fmicb.2018.02564.
- Guo Y., Han M., Xu J., Zhang W. Analysis of acetohydroxyacid synthase variants from branched-chain amino acids-producing strains and their effects on the synthesis of branched-chain amino acids in *Corynebacterium glutamicum*. *Protein Expr. Purif.* 2015;109:106-112. DOI 10.1016/j.pep.2015.02.006.
- Guo Y., Han M., Yan W., Xu J., Zhang W. Generation of branched-chain amino acids resistant *Corynebacterium glutamicum* acetohydroxy acid synthase by site-directed mutagenesis. *Biotechnol. Bioproc. Eng.* 2014;19:456-467. DOI 10.1007/s12257-013-0843-x.
- Han G., Xu N., Sun X., Chen J., Chen C., Wang Q. Improvement of L-valine production by atmospheric and room temperature plasma mutagenesis and high-throughput screening in *Corynebacterium glutamicum*. *ACS Omega*. 2020;5(10):4751-4758. DOI 10.1021/acsomega.9b02747.
- Harst A., Albaum S.P., Bojarzyn T., Trötschel C., Poetsch A. Proteomics of FACS-sorted heterogeneous *Corynebacterium glutamicum* populations. *J. Proteomics*. 2017;160:1-7. DOI 10.1016/j.jprot.2017.03.01.
- Hasegawa S., Suda M., Uematsu K., Natsuma Y., Hiraga K., Jojima T., Inui M., Yukawa H. Engineering of *Corynebacterium glutamicum* for high-yield L-valine production under oxygen deprivation conditions. *Appl. Environ. Microbiol.* 2013;79(4):1250-1257. DOI 10.1128/aem.02806-12.
- Hasegawa S., Uematsu K., Natsuma Y., Suda M., Hiraga K., Jojima T., Inui M., Yukawa H. Improvement of the redox balance increases L-valine production by *Corynebacterium glutamicum* under oxygen deprivation conditions. *Appl. Environ. Microbiol.* 2012;78(3):865-875. DOI 10.1128/aem.07056-11.
- Hemmerich J., Tenhaef N., Steffens C., Kappelmann J., Weiske M., Reich S.J., Wiechert W., Oldiges M., Noack S. Less sacrifice, more insight: Repeated low-volume sampling of microbioreactor cultivations enables accelerated deep phenotyping of microbial strain libraries. *Biotechnol. J.* 2018;14(9):e1800428. DOI 10.1002/biot.201800428.
- Hermann T., Kramer R. Mechanism and regulation of isoleucine excretion in *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 1996;62(9):3238-3244. DOI 10.1128/aem.62.9.3238-3244.1996.
- Holátko J., Elišáková V., Prouza M., Sobotka M., Nesvera J., Pátek M. Metabolic engineering of the L-valine biosynthesis pathway in *Corynebacterium glutamicum* using promoter activity modulation. *J. Biotechnol.* 2009;139(3):203-210. DOI 10.1016/j.jbiotec.2008.12.005.
- Holeček M. Branched-chain amino acids in health and disease: metabolism, alterations in blood plasma, and as supplements. *Nutr. Metab. (Lond)*. 2018;15:33. DOI 10.1186/s12986-018-0271-1.
- Holen J.P., Tokach M.D., Woodworth J.C., DeRouchey J.M., Gebhardt J.T., Titgemeyer E.C., Goodband R.D. A review of branched-chain amino acids in lactation diets on sow and litter growth performance. *Transl. Anim. Sci.* 2022;6(1):txac017. DOI 10.1093/tas/txac017.
- Hou X.H., Chen X.D., Zhang Y., Qian H., Zhang W.G. L-valine production with minimization of by-products' synthesis in *Corynebacterium glutamicum* and *Brevibacterium flavum*. *Amino Acids*. 2012a;43(6):2301-2311. DOI 10.1007/s00726-012-1308-9.
- Hou X.H., Ge X.Y., Wu D., Qian H., Zhang W.G. Improvement of L-valine production at high temperature in *Brevibacterium flavum* by overexpressing *ilvEBN(r)C* genes. *J. Ind. Microbiol. Biotechnol.* 2012b;39(1):63-72. DOI 10.1007/s10295-011-1000-1.
- Jian H., Miao S., Liu Y., Li H., Zhou W., Wang X., Dong X., Zou X. Effects of dietary valine levels on production performance, egg quality, antioxidant capacity, immunity, and intestinal amino acid absorption of laying hens during the peak lay period. *Animals (Basel)*. 2021;11(7):1972. DOI 10.3390/ani11071972.
- Jiang L.Y., Zhang Y.Y., Li Z., Liu J.Z. Metabolic engineering of *Corynebacterium glutamicum* for increasing the production of L-ornithine by increasing NADPH availability. *J. Ind. Microbiol. Biotechnol.* 2013;40(10):1143-1151. DOI 10.1007/s10295-013-1306-2.
- Jojima T., Fujii M., Mori E., Inui M., Yukawa H. Engineering of sugar metabolism of *Corynebacterium glutamicum* for production of amino acid L-alanine under oxygen deprivation. *Appl. Microbiol. Biotechnol.* 2010;87(1):159-165. DOI 10.1007/s00253-010-2493-7.
- Jojima T., Noburyu R., Sasaki M., Tajima T., Suda M., Yukawa H., Inui M. Metabolic engineering for improved production of ethanol by *Corynebacterium glutamicum*. *Appl. Microbiol. Biotechnol.* 2015;99(3):1165-1172. DOI 10.1007/s00253-014-6223-4.
- Jones C.M., Hernandez Lozada N.J., Pfleger B.F. Efflux systems in bacteria and their metabolic engineering applications. *Appl. Microbiol. Biotechnol.* 2015;99(22):9381-9393. DOI 10.1007/s00253-015-6963-9.

- Kabus A., Georgi T., Wendisch V.F., Bott M. Expression of the *Escherichia coli pntAB* genes encoding a membrane-bound transhydrogenase in *Corynebacterium glutamicum* improves L-lysine formation. *Appl. Microbiol. Biotechnol.* 2007;75(1):47-53. DOI 10.1007/s00253-006-0804-9.
- Kainulainen H., Hulmi J.J., Kujala U.M. Potential role of branched-chain amino acid catabolism in regulating fat oxidation. *Exerc. Sport Sci. Rev.* 2013;41(4):194-200. DOI 10.1097/JES.0b013e3182a4e6b6.
- Kang K.Y., Kim M.S., Lee M.S., Oh J.J., An S., Park D., Heo I.K., Lee H.K., Song S.W., Kim S.D. Genotoxicity and acute toxicity evaluation of the three amino acid additives with *Corynebacterium glutamicum* biomass. *Toxicol. Rep.* 2020;7:241-253. DOI 10.1016/j.toxrep.2020.01.013.
- Karau A., Grayson I. Amino acids in human and animal nutrition. *Adv. Biochem. Eng. Biotechnol.* 2014;143:189-228. DOI 10.1007/10_2014_269.
- Kataoka N., Vangnai A.S., Pongtharangkul T., Yakushi T., Wada M., Yokota A., Matsushita K. Engineering of *Corynebacterium glutamicum* as a prototrophic pyruvate-producing strain: Characterization of a *ramA*-deficient mutant and its application for metabolic engineering. *Biosci. Biotechnol. Biochem.* 2019;83(2):372-380. DOI 10.1080/09168451.2018.1527211.
- Kawaguchi T., Izumi N., Charlton M.R., Sata M. Branched-chain amino acids as pharmacological nutrients in chronic liver disease. *Hepatology*. 2011;54(3):1063-1070. DOI 10.1002/hep.24412.
- Keilhauer C., Eggeling L., Sahm H. Isoleucine synthesis in *Corynebacterium glutamicum*: molecular analysis of the *ilvB-ilvN-ilvC* operon. *J. Bacteriol.* 1993;175(17):5595-5603. DOI 10.1128/jb.175.17.5595-5603.1993.
- Kennerknecht N., Sahm H., Yen M.R., Pátek M., Saier M.H. Jr., Eggeling L. Export of L-isoleucine from *Corynebacterium glutamicum*: a two-gene-encoded member of a new translocator family. *J. Bacteriol.* 2002;184(14):3947-3956. DOI 10.1128/jb.184.14.3947-3956.2002.
- Koduru L., Lakshmanan M., Lee D.Y. In silico model-guided identification of transcriptional regulator targets for efficient strain design. *Microb. Cell Fact.* 2018;17(1):167. DOI 10.1186/s12934-018-1015-7.
- Krause F.S., Blombach B., Eikmanns B.J. Metabolic engineering of *Corynebacterium glutamicum* for 2-ketoisovalerate production. *Appl. Environ. Microbiol.* 2010a;76(24):8053-8061. DOI 10.1128/aem.01710-10.
- Krause F.S., Henrich A., Blombach B., Kramer R., Eikmanns B.J., Seibold G.M. Increased glucose utilization in *Corynebacterium glutamicum* by use of maltose, and its application for the improvement of L-valine productivity. *Appl. Environ. Microbiol.* 2010b;76(1):370-374. DOI 10.1128/aem.01553-09.
- Lange C., Mustafi N., Frunzke J., Kennerknecht N., Wessel M., Bott M., Wendisch V.F. Lrp of *Corynebacterium glutamicum* controls expression of the *brnFE* operon encoding the export system for L-methionine and branched-chain amino acids. *J. Biotechnol.* 2012;158(4):231-241. DOI 10.1016/j.jbiotec.2011.06.003.
- Lange J., Münch E., Müller J., Busche T., Kalinowski J., Takors R., Blombach B. Deciphering the adaptation of *Corynebacterium glutamicum* in transition from aerobiosis via microaerobiosis to anaerobiosis. *Genes (Basel)*. 2018;9(6):297. DOI 10.3390/genes9060297.
- Lee D., Hong J., Kim K.J. Crystal structure and biochemical characterization of ketol-acid reductoisomerase from *Corynebacterium glutamicum*. *J. Agric. Food Chem.* 2019;67(31):8527-8535. DOI 10.1021/acs.jafc.9b03262.
- Leuchtenberger W., Huthmacher K., Drauz K. Biotechnological production of amino acids and derivatives: current status and prospects. *Appl. Microbiol. Biotechnol.* 2005;69(1):1-8. DOI 10.1007/s00253-005-0155-y.
- Leyval D., Uy D., Delaunay S., Goergen J.L., Engasser J.M. Characterisation of the enzyme activities involved in the valine biosynthetic pathway in a valine-producing strain of *Corynebacterium glutamicum*. *J. Biotechnol.* 2003;104(1-3):241-252. DOI 10.1016/s0168-1656(03)00162-7.
- Li N., Xu S., Du G., Chen J., Zhou J. Efficient production of L-homoserine in *Corynebacterium glutamicum* ATCC 13032 by redistribution of metabolic flux. *Biochem. Eng. J.* 2020a;161:107665. DOI 10.1016/j.bej.2020.107665.
- Li N., Zeng W., Xu S., Zhou J. Obtaining a series of native gradient promoter-5'-UTR sequences in *Corynebacterium glutamicum* ATCC 13032. *Microb. Cell. Fact.* 2020b;19(1):120. DOI 10.1186/s12934-020-01376-3.
- Li Y., Cong H., Liu B., Song J., Sun X., Zhang J., Yang Q. Metabolic engineering of *Corynebacterium glutamicum* for methionine production by removing feedback inhibition and increasing NADPH level. *Antonie Van Leeuwenhoek*. 2016;109(9):1185-1197. DOI 10.1007/s10482-016-0719-0.
- Lindner S.N., Petrov D.P., Hagmann C.T., Henrich A., Krämer R., Eikmanns B.J., Wendisch V.F., Seibold G.M. Phosphotransferase system-mediated glucose uptake is repressed in phosphoglucosomerase-deficient *Corynebacterium glutamicum* strains. *Appl. Environ. Microbiol.* 2013;79(8):2588-2595. DOI 10.1128/AEM.03231-12.
- Liu Y., Li Y., Wang X. Acetohydroxyacid synthases: evolution, structure, and function. *Appl. Microbiol. Biotechnol.* 2016;100(20):8633-8649. DOI 10.1007/s00253-016-7809-9.
- Liu Y., Wang X., Zhan J., Hu J. The 138th residue of acetohydroxyacid synthase in *Corynebacterium glutamicum* is important for the substrate binding specificity. *Enzyme Microb. Technol.* 2019;129:109357. DOI 10.1016/j.enzmictec.2019.06.001.
- Liu Y., Zhang C., Zhang Y., Jiang X., Liang Y., Wang H., Li Y., Sun G. Association between excessive dietary branched-chain amino acids intake and hypertension risk in chinese population. *Nutrients*. 2022;14(13):2582. DOI 10.3390/nu14132582.
- Ma Y., Chen Q., Cui Y., Du L., Shi T., Xu Q., Ma Q., Xie X., Chen N. Comparative genomic and genetic functional analysis of industrial L-leucine- and L-valine-producing *Corynebacterium glutamicum* strains. *J. Microbiol. Biotechnol.* 2018a;28(11):1916-1927. DOI 10.4014/jmb.1805.05013.
- Ma Y., Cui Y., Du L., Liu X., Xie X., Chen N. Identification and application of a growth-regulated promoter for improving L-valine production in *Corynebacterium glutamicum*. *Microb. Cell. Fact.* 2018b;17(1):185. DOI 10.1186/s12934-018-1031-7.
- Magnus J.B., Oldiges M., Takors R. The identification of enzyme targets for the optimization of a valine producing *Corynebacterium glutamicum* strain using a kinetic model. *Biotechnol. Prog.* 2009;25(3):754-762. DOI 10.1002/btpr.184.
- Marienhagen J., Eggeling L. Metabolic function of *Corynebacterium glutamicum* aminotransferases AlaT and AvtA and impact on L-valine production. *Appl. Environ. Microbiol.* 2008;74(24):7457-7462. DOI 10.1128/AEM.01025-08.
- Marienhagen J., Kennerknecht N., Sahm H., Eggeling L. Functional analysis of all aminotransferase proteins inferred from the genome sequence of *Corynebacterium glutamicum*. *J. Bacteriol.* 2005;187(22):7639-7646. DOI 10.1128/JB.187.22.7639-7646.2005.
- Marx A., Striegel K., de Graaf A.A., Sahm H., Eggeling L. Response of the central metabolism of *Corynebacterium glutamicum* to different flux burdens. *Biotechnol. Bioeng.* 1997;56(2):168-180. DOI 10.1002/(SICI)1097-0290(19971020)56:2<168::AID-BIT6>3.0.CO;2-N.
- Michel A., Koch-Koerfges A., Krumbach K., Brocker M., Bott M. Anaerobic growth of *Corynebacterium glutamicum* via mixed-acid fermentation. *Appl. Environ. Microbiol.* 2015;81(21):7496-7508. DOI 10.1128/AEM.02413-15.
- Möckel B., Eggeling L., Sahm H. Functional and structural analyses of threonine dehydratase from *Corynebacterium glutamicum*. *J. Bacteriol.* 1992;174(24):8065-8072. DOI 10.1128/jb.174.24.8065-8072.1992.
- Morbach S., Junger C., Sahm H., Eggeling L. Attenuation control of *ilvBNC* in *Corynebacterium glutamicum*: evidence of leader peptide

- formation without the presence of a ribosome binding site. *J. Biosci. Bioeng.* 2000;90(5):501-507. DOI 10.1016/S1389-1723(01)80030-X.
- Moritz B., Striegel K., De Graaf A.A., Sahm H. Kinetic properties of the glucose-6-phosphate and 6-phosphogluconate dehydrogenases from *Corynebacterium glutamicum* and their application for predicting pentose phosphate pathway flux *in vivo*. *Eur. J. Biochem.* 2000;267(12):3442-3452. DOI 10.1046/j.1432-1327.2000.01354.x.
- Okino S., Suda M., Fujikura K., Inui M., Yukawa H. Production of D-lactic acid by *Corynebacterium glutamicum* under oxygen deprivation. *Appl. Microbiol. Biotechnol.* 2008;78(3):449-454. DOI 10.1007/s00253-007-1336-7.
- Park J.H., Lee S.Y. Fermentative production of branched chain amino acids: a focus on metabolic engineering. *Appl. Microbiol. Biotechnol.* 2010;85(3):491-506. DOI 10.1007/s00253-009-2307-y.
- Pérez-García F., Jorge J.M.P., Dreyszas A., Risse J.M., Wendisch V.F. Efficient production of the dicarboxylic acid glutarate by *Corynebacterium glutamicum* via a novel synthetic pathway. *Front. Microbiol.* 2018;9:2589. DOI 10.3389/fmicb.2018.02589.
- Pérez-García F., Wendisch V.F. Transport and metabolic engineering of the cell factory *Corynebacterium glutamicum*. *FEMS Microbiol. Lett.* 2018;365(16):fny166. DOI 10.1093/femsle/fny166.
- Qin T., Hu X., Hu J., Wang X. Metabolic engineering of *Corynebacterium glutamicum* strain ATCC13032 to produce L-methionine. *Biotechnol. Appl. Biochem.* 2015;62(4):563-673. DOI 10.1002/bab.1290.
- Radmacher E., Vaitsikova A., Burger U., Krumbach K., Sahm H., Eggeling L. Linking central metabolism with increased pathway flux: L-valine accumulation by *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 2002;68(5):2246-2250. DOI 10.1128/aem.68.5.2246-2250.2002.
- Ruklisha M., Paegle L., Denina I. L-Valine biosynthesis during batch and fed-batch cultivations of *Corynebacterium glutamicum*: Relationship between changes in bacterial growth rate and intracellular metabolism. *Proc. Biochem.* 2007;40(4):634-640. DOI 10.1016/j.procbio.2006.11.008.
- Ryabchenko L.E., Gerasimova T.V., Leonova T.E., Kalinina T.I., Shermetyeva M.E., Anufriev K.E., Yanenko A.S. Patent RU 2753996 C1. Bacterium *Corynebacterium glutamicum* with increased ability to produce L-valine and method for producing L-valine using this bacterium. Date of publication: 25.08.2021. Bull. No. 24. (in Russian)
- Sahm H., Eggeling L. D-pantothenate synthesis in *Corynebacterium glutamicum* and use of *panBC* and genes encoding L-valine synthesis for D-pantothenate overproduction. *Appl. Environ. Microbiol.* 1999;65(5):1973-1979. DOI 10.1128/AEM.65.5.1973-1979.1999.
- Savrasova E.A., Stoyanova N.V. Application of leucine dehydrogenase Bcd from *Bacillus subtilis* for L-valine synthesis in *Escherichia coli* under microaerobic conditions. *Heliyon.* 2019;5(4):e01406. DOI 10.1016/j.heliyon.2019.e01406.
- Schwentner A., Feith A., Münch E., Busche T., Rückert C., Kalinowski J., Takors R., Blombach B. Metabolic engineering to guide evolution – Creating a novel mode for L-valine production with *Corynebacterium glutamicum*. *Metab. Eng.* 2018;47:31-41. DOI 10.1016/j.ymben.2018.02.015.
- Shi F., Li K., Huan X., Wang X. Expression of NAD(H) kinase and glucose-6-phosphate dehydrogenase improve NADPH supply and L-isoleucine biosynthesis in *Corynebacterium glutamicum* ssp. *lactofermentum*. *Appl. Biochem. Biotechnol.* 2013;171(2):504-521. DOI 10.1007/s12010-013-0389-6.
- Shi F., Luan M., Li Y. Ribosomal binding site sequences and promoters for expressing glutamate decarboxylase and producing γ -aminobutyrate in *Corynebacterium glutamicum*. *AMB Express.* 2018; 8(1):61. DOI 10.1186/s13568-018-0595-2.
- Shou J., Chen P.J., Xiao W.H. The effects of BCAAs on insulin resistance in athletes. *J. Nutr. Sci. Vitaminol. (Tokyo).* 2019;65(5):383-389. DOI 10.3177/jnsv.65.383.
- Siedler S., Lindner S.N., Bringer S., Wendisch V.F., Bott M. Reductive whole-cell biotransformation with *Corynebacterium glutamicum*: improvement of NADPH generation from glucose by a cyclized pentose phosphate pathway using *pfkA* and *gapA* deletion. *Appl. Microbiol. Biotechnol.* 2013;97(1):143-152. DOI 10.1007/s00253-012-4314-7.
- Tarutina M.G., Raevskaya N.M., Shustikova T.E., Ryabchenko L.E., Yanenko A.S. Assessment of effectiveness of *Corynebacterium glutamicum* promoters and their application for the enhancement of gene activity in lysine-producing bacteria. *Appl. Biochem. Microbiol.* 2016;52(7):692-698. DOI 10.1134/S0003683816070073.
- Tauch A., Hermann T., Burkovski A., Kramer R., Puhler A., Kalinowski J. Isoleucine uptake in *Corynebacterium glutamicum* ATCC 13032 is directed by the *brnQ* gene product. *Arch. Microbiol.* 1998;169(4):303-312. DOI 10.1007/s002030050576.
- Trotschel C., Deutenberg D., Bathe B., Burkovski A., Kramer R. Characterization of methionine export in *Corynebacterium glutamicum*. *J. Bacteriol.* 2005;187(11):3786-3794. DOI 10.1128/jb.187.11.3786-3794.2005.
- Vasicová P., Pátek M., Nesvera J., Sahm H., Eikmanns B. Analysis of the *Corynebacterium glutamicum* *dapA* promoter. *J. Bacteriol.* 1999; 181(19):6188-6191. DOI 10.1128/JB.181.19.6188-6191.1999.
- Vogt M., Haas S., Klaffl S., Polen T., Eggeling L., van Ooyen J., Bott M. Pushing product formation to its limit: metabolic engineering of *Corynebacterium glutamicum* for L-leucine overproduction. *Metab. Eng.* 2014;22:40-52. DOI 10.1016/j.ymben.2013.12.001.
- Wang X., Zhang H., Quinn P.J. Production of L-valine from metabolically engineered *Corynebacterium glutamicum*. *Appl. Microbiol. Biotechnol.* 2018;102(10):4319-4330. DOI 10.1007/s00253-018-8952-2.
- Wang Y.Y., Shi K., Chen P., Zhang F., Xu J.Z., Zhang W.G. Rational modification of the carbon metabolism of *Corynebacterium glutamicum* to enhance L-leucine production. *J. Ind. Microbiol. Biotechnol.* 2020;47(6-7):485-495. DOI 10.1007/s10295-020-02282-8.
- Wang Y.Y., Xu J.Z., Zhang W.G. Metabolic engineering of L-leucine production in *Escherichia coli* and *Corynebacterium glutamicum*: a review. *Crit. Rev. Biotechnol.* 2019a;39(5):633-647. DOI 10.1080/07388551.2019.1577214.
- Wang Y.Y., Zhang F., Xu J.Z., Zhang W.G., Chen X.L., Liu L.M. Improvement of L-leucine production in *Corynebacterium glutamicum* by altering the redox flux. *Int. J. Mol. Sci.* 2019b;20(8):2020. DOI 10.3390/ijms20082020.
- Wang Z., Chen T., Ma X., Shen Z., Zhao X. Enhancement of riboflavin production with *Bacillus subtilis* by expression and site-directed mutagenesis of *zwf* and *gnd* gene from *Corynebacterium glutamicum*. *Bioresour. Technol.* 2011;102(4):3934-3940. DOI 10.1016/j.biortech.2010.11.120.
- Wei H., Ma Y., Chen Q., Cui Y., Du L., Ma Q., Li Y., Xie X., Chen N. Identification and application of a novel strong constitutive promoter in *Corynebacterium glutamicum*. *Ann. Microbiol.* 2018;68:375-382. DOI 10.1007/s13213-018-1344-0.
- Wieschalka S., Blombach B., Bott M., Eikmanns B.J. Bio-based production of organic acids with *Corynebacterium glutamicum*. *Microb. Biotechnol.* 2012;6(2):87-102. DOI 10.1111/1751-7915.12013.
- Xie X., Xu L., Shi J., Xu Q., Chen N. Effect of transport proteins on L-isoleucine production with the L-isoleucine-producing strain *Corynebacterium glutamicum* YILW. *J. Ind. Microbiol. Biotechnol.* 2012;39(10):1549-1556. DOI 10.1007/s10295-012-1155-4.
- Xu J., Han M., Zhang J., Guo Y., Zhang W. Metabolic engineering *Corynebacterium glutamicum* for the L-lysine production by increasing the flux into L-lysine biosynthetic pathway. *Amino Acids.* 2014;46(9):2165-2175. DOI 10.1007/s00726-014-1768-1.
- Xu J.Z., Yu H.B., Han M., Liu L.M., Zhang W.G. Metabolic engineering of glucose uptake systems in *Corynebacterium glutamicum* for improving the efficiency of L-lysine production. *J. Ind. Microbiol. Biotechnol.* 2019;46(7):937-949. DOI 10.1007/s10295-019-02170-w.
- Xu N., Wei L., Liu J. Recent advances in the applications of promoter engineering for the optimization of metabolite biosynthesis. *World*

- J. Microbiol. Biotechnol.* 2019;35(2):33. DOI 10.1007/s11274-019-2606-0.
- Yamamoto K., Tsuchisaka A., Yukawa H. Branched-chain amino acids. *Adv. Biochem. Eng. Biotechnol.* 2017;159:103-128. DOI 10.1007/10_2016_28.
- Yamamoto S., Suda M., Niimi S., Inui M., Yukawa H. Strain optimization for efficient isobutanol production using *Corynebacterium glutamicum* under oxygen deprivation. *Biotechnol. Bioeng.* 2013;110(11):2938-2948. DOI 10.1002/bit.24961.
- Yin L., Shi F., Hu X., Chen C., Wang X. Increasing L-isoleucine production in *Corynebacterium glutamicum* by overexpressing global regulator Lrp and two-component export system BrnFE. *J. Appl. Microbiol.* 2013;114(5):1369-1377. DOI 10.1111/jam.12141.
- Yin L., Zhao J., Chen C., Xu X., Wang X. Enhancing the carbon flux and NADPH supply to increase L-isoleucine production in *Corynebacterium glutamicum*. *Biotechnol. Bioproc. Eng.* 2014;19:132-142. DOI 10.1007/s12257-013-0416-z.
- Zhan M., Kan B., Dong J., Xu G., Han R., Ni Y. Metabolic engineering of *Corynebacterium glutamicum* for improved L-arginine synthesis by enhancing NADPH supply. *J. Ind. Microbiol. Biotechnol.* 2019;46(1):45-54. DOI 10.1007/s10295-018-2103-8.
- Zhang H., Li Y., Wang C., Wang X. Understanding the high L-valine production in *Corynebacterium glutamicum* VWB-1 using transcriptomics and proteomics. *Sci. Rep.* 2018;8(1):3632. DOI 10.1038/s41598-018-21926-5.
- Zhang J., Qian F., Dong F., Wang Q., Yang J., Jiang Y., Yang S. *De novo* engineering of *Corynebacterium glutamicum* for L-proline production. *ACS Synth. Biol.* 2020;9(7):1897-1906. DOI 10.1021/acssynbio.0c00249.
- Zhang S., Liu D., Mao Z., Mao Y., Ma H., Chen T., Zhao X., Wang Z. Model-based reconstruction of synthetic promoter library in *Corynebacterium glutamicum*. *Biotechnol. Lett.* 2018;40(5):819-827. DOI 10.1007/s10529-018-2539-y.
- Zhang Y., Liu Y., Zhang S., Ma W., Wang J., Yin L., Wang X. Metabolic engineering of *Corynebacterium glutamicum* WM001 to improve L-isoleucine production. *Biotechnol. Appl. Biochem.* 2021;68(3):568-584. DOI 10.1002/bab.1963.
- Zheng L., Zuo F., Zhao S., He P., Wei H., Xiang Q., Pang J., Peng J. Dietary supplementation of branched-chain amino acids increases muscle net amino acid fluxes through elevating their substrate availability and intramuscular catabolism in young pigs. *Br. J. Nutr.* 2017;117(7):911-922. DOI 10.1017/S0007114517000757.

Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Projects No. 075-15-2019-1659 and 075-15-2019-1662).

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

Received August 7, 2022. Revised October 26, 2022. Accepted October 26, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Stratifications and foliations in phase portraits of gene network models

V.P. Golubyatnikov^{1, 3}✉, A.A. Akinshin², N.B. Ayupova^{1, 3}, L.S. Minushkina³

¹ Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Huawei Russian Research Institute, St. Petersburg, Russia

³ Novosibirsk State University, Novosibirsk, Russia

✉ golubyatn@yandex.ru

Abstract. Periodic processes of gene network functioning are described with good precision by periodic trajectories (limit cycles) of multidimensional systems of kinetic-type differential equations. In the literature, such systems are often called dynamical, they are composed according to schemes of positive and negative feedback between components of these networks. The variables in these equations describe concentrations of these components as functions of time. In the preparation of numerical experiments with such mathematical models, it is useful to start with studies of qualitative behavior of ensembles of trajectories of the corresponding dynamical systems, in particular, to estimate the highest likelihood domain of the initial data, to solve inverse problems of parameter identification, to list the equilibrium points and their characteristics, to localize cycles in the phase portraits, to construct stratification of the phase portraits to subdomains with different qualities of trajectory behavior, etc. Such an *à priori* geometric analysis of the dynamical systems is quite analogous to the basic section “Investigation of functions and plot of their graphs” of Calculus, where the methods of qualitative studies of shapes of curves determined by equations are exposed. In the present paper, we construct ensembles of trajectories in phase portraits of some dynamical systems. These ensembles are 2-dimensional surfaces invariant with respect to shifts along the trajectories. This is analogous to classical construction in analytic mechanics, i.e. the level surfaces of motion integrals (energy, kinetic moment, etc.). Such surfaces compose foliations in phase portraits of dynamical systems of Hamiltonian mechanics. In contrast with this classical mechanical case, the foliations considered in this paper have singularities: all their leaves have a non-empty intersection, they contain limit cycles on their boundaries. Description of the phase portraits of these systems at the level of their stratifications, and that of ensembles of trajectories allows one to construct more realistic gene network models on the basis of methods of statistical physics and the theory of stochastic differential equations.

Key words: oscillations; positive and negative feedbacks; gene network models; phase portraits; invariant domains and surfaces; invariant foliations; Poincaré map; Grobman–Hartman theorem; Frobenius–Perron theorem.

For citation: Golubyatnikov V.P., Akinshin A.A., Ayupova N.B., Minushkina L.S. Stratifications and foliations in phase portraits of gene network models. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):758-764. DOI 10.18699/VJGB-22-91

Стратификации и слоения в фазовых портретах моделей генных сетей

В.П. Голубятников^{1, 3}✉, А.А. Акиншин², Н.Б. Аюпова^{1, 3}, Л.С. Минушкина³

¹ Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия

² Российский исследовательский институт Huawei, Санкт-Петербург, Россия

³ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ golubyatn@yandex.ru

Аннотация. Периодические процессы функционирования широкого класса генных сетей с хорошей точностью описываются предельными циклами многомерных систем дифференциальных уравнений кинетического типа. Такие системы, часто называемые в литературе динамическими, состояются по схемам положительных и отрицательных связей между компонентами моделируемых сетей. Искомые функции в уравнениях описывают зависимость от времени концентраций этих компонент. При планировании вычислительных экспериментов с подобными математическими моделями полезно предварительно описать качественное поведение ансамблей траекторий соответствующих динамических систем, в частности оценить области максимального правдоподобия начальных данных, исследовать обратные задачи идентификации параметров, особые точки этих систем, локализовать в фазовых портретах положение циклов, в том числе предельных,

стратифицировать фазовые портреты на подобласти с качественно различным поведением траекторий и т.п. Такой априорный геометрический анализ рассматриваемых моделей генных сетей полностью аналогичен хрестоматийному разделу начальных курсов математики «Исследование функций и построение графиков», в котором описываются методы наглядного представления поведения кривых, определяемых уравнениями. В настоящей статье в фазовых портретах динамических систем, моделирующих функционирование кольцевых генных сетей, конструируются двумерные поверхности, инвариантные относительно сдвигов вдоль траекторий, – ансамбли траекторий. Просматривается естественная аналогия с классической конструкцией аналитической механики – с поверхностями уровня интегралов движения (энергия, импульс и др.). Такие поверхности образуют слоения в фазовых портретах динамических систем гамильтоновой механики. В отличие от задач механики, для рассматриваемых нами моделей генных сетей слоения, образуемые инвариантными поверхностями, имеют особенности, все их слои содержат на своих границах предельные циклы. Описание фазовых портретов динамических систем в терминах их стратификаций и ансамблей их траекторий позволит строить более реалистичные модели генных сетей с использованием аппарата статистической физики и теории стохастических дифференциальных уравнений.

Ключевые слова: осцилляции; положительные и отрицательные связи; модели генных сетей; фазовые портреты; инвариантные области и поверхности; инвариантные слоения; отображение Пуанкаре; теорема Гробмана–Хартмана; теорема Фробениуса–Перрона.

Introduction

At present time, investigation of questions of existence of periodic trajectories (cycles) in phase portraits of systems of non-linear differential equations simulating functioning of various natural processes is carried out in most fields of applied mathematics. Detection of such cycles, their localization in the phase portraits, description of their characteristics, such as stability, (non)uniqueness, etc. have a long history (Poincaré, 1892). These problems have generated a whole range of research directions in pure mathematics: qualitative theory of differential equations, theory of dynamics systems, etc., which in turn have a great impact on corresponding applied disciplines. At their junction, the famous 16-th Hilbert's problem, and the “center-focus” problem, related to seemingly just a pictorial case of two differential equations with two unknown functions of one variable (time) have appeared.

Here, in the present paper, we study systems of kinetic equations of higher dimensions, considered as functioning of circular gene networks models:

$$\frac{dx_j}{dt} = f_j(x_{j-1}) - k_j x_j. \quad (1)$$

It is assumed here and below that $j = 1, 2, \dots, n$; $n \geq 3$, and that $j-1 = n$, if $j = 1$. In all these equations, non-negative functions $x_j(t)$ denote concentrations of species in the gene networks, and positive coefficients k_j characterize the rates of their degradations (Likhoshvai et al., 2020).

Consider the system (1) in the vector form $\frac{dX}{dt} = F(X)$, where the vector-function $X(t)$ is defined by its coordinate functions $x_j(t)$. The divergence of this vector-field $F(X)$ is constant and negative:

$$\operatorname{div} F(X) \equiv -k_1 - k_2 - \dots - k_n < 0.$$

It is well-known (Arnold, 1989) that in this case, n -dimensional volume of any finite domain in the phase portrait decreases exponentially during the shifts of its points along trajectories of the system (1) as t grows. This does not mean that each such domain collapses to a point. For the dynamical systems considered here, these limit sets are two-dimensional invariant surfaces in their n -dimensional phase portraits.

We call the dynamical system (1) block-linear if for all j each function f_j which describes the rate of synthesis of the j -th component of the gene network is a step-function (threshold function)

$$f_j(y) \equiv L_j(y) = k_j a_j, \text{ if } 0 \leq y \leq 1; L_j(y) \equiv 0, \text{ if } y > 1; \\ \text{or } f_j(y) \equiv \Gamma_j(y) \equiv 0, \text{ if } 0 \leq y \leq 1; \Gamma_j(y) \equiv k_j a_j, \text{ if } y > 1.$$

Here, a_j are some positive constants. Decreasing functions L_j describe negative feedbacks in the gene network and increasing functions Γ_j correspond to positive feedbacks.

For one particular case $k_j = 1$ for all j , investigation of cycles of similar block-linear systems was realized in (Glass, Pasternack, 1978; Akinshin et al., 2013; Ayupova, Golubyatnikov, 2014; Golubyatnikov, Gradov, 2021). Under the same assumptions, questions of existence of cycles in smooth analogues of these systems were studied in (Elowitz, Leibler, 2000; Glyzin et al., 2016; Kolesov et al., 2016) in the cases when these systems are symmetric with respect to cyclic permutations of pairs of the variables x_j .

In recent publications (Golubyatnikov, Ivanov, 2018; Golubyatnikov, Minushkina, 2019, 2020; Likhoshvai et al., 2020; Ivanov, 2022), existence, uniqueness, and stability of the cycles of block-linear dynamical systems of some different dimensions with arbitrary positive coefficients k_j were proved with the help of stratification of phase portraits to subdomains according to behavior of trajectories. It was shown there that these phase portraits contain cycles if and only if $a_j > 1$ for all j and that the parallelepiped $Q^n = [0, a_1] \times [0, a_2] \times \dots \times [0, a_n]$ in the positive octant of the space \mathbf{R}^n is a positively invariant domain of the dynamical system (1). This means that trajectories of all points of the domain Q^n do not leave it and that all cycles of the system (1) are contained in the interior of Q^n . We consider below the dynamical systems of the type (1) in the case $a_j > 1$ for all j only. Physical interpretation of this condition means that the maximal rate of synthesis of any component of the gene network exceeds that of its degradation.

We decompose the domain Q^n by the planes $x_j = 1$ to 2^n smaller parallelepipeds, which we call blocks and enumerate by binary multi-indices: $\{\varepsilon_1 \varepsilon_2 \dots \varepsilon_n\} = I_1(\varepsilon_1) \times I_2(\varepsilon_2) \times \dots \times I_n(\varepsilon_n)$.

Here, each index ε_j equals 0 or 1, and $I_j(0) = [0, 1]$, $I_j(1) = (1, a_j]$. Let E be the common point of all these blocks (all its coordinates equal one). In each of these blocks, the equations of the system (1) take the simplest linear form

$$\frac{dx_j}{dt} = k_j(x_j - a_j(1 - \varepsilon_{j-1})),$$

and solution to the Cauchy problem for this system has a simple representation

$$x_j(t) = a_j(1 - \varepsilon_{j-1}) + (x_j(0) - a_j(1 - \varepsilon_{j-1})) \exp(-k_j t). \quad (2)$$

In the present paper, for some low dimensional block-linear dynamical systems considered as models of gene networks functioning, we study the behavior of ensembles of their trajectories and show the existence of families of two-dimensional surfaces that are invariant with respect to shifts along trajectories of these systems and contain their cycles. This makes the qualitative analysis of trajectory behavior and interpretation of numerical experiments with these models much simpler.

Three-dimensional dynamical system

In the papers (Golubyatnikov et al., 2018; Golubyatnikov, Ivanov, 2018), we considered a 3D block-linear dynamic system:

$$\frac{dx_1}{dt} = L_1(x_3) - k_1 x_1; \quad \frac{dx_2}{dt} = L_2(x_1) - k_2 x_2; \quad \frac{dx_3}{dt} = L_3(x_2) - k_3 x_3. \quad (3)$$

Trajectories of all points of the block $\{001\}$ pass through six blocks of decomposition of the domain Q^3 from block to block according to arrows of the following diagram only:

$$\begin{aligned} & \dots \rightarrow \{001\} \rightarrow \{011\} \rightarrow \{010\} \rightarrow \\ & \{110\} \rightarrow \{100\} \rightarrow \{101\} \rightarrow \{001\} \dots \end{aligned} \quad (4)$$

Denote by W_1^3 a union of blocks listed in the diagram, this is a positive invariant domain of the system (3), its interior is homeomorphic to torus. Note that trajectories of points of two blocks, $\{000\}$ and $\{111\}$, eventually leave them in the invariant domain W_1^3 and further stay there. Thus, cycles of the system (3) do not intersect these two blocks (Golubyatnikov et al., 2018). Stratification of phase portrait of the system (3) consists of two parts: the domain W_1^3 and the union of two blocks, $\{000\}$, $\{111\}$.

Consider a two-dimensional face $F_0 = \{001\} \cap \{011\}$ which separates the blocks $\{001\}$ and $\{011\}$ as well as other faces F_m which separate incident blocks of the diagram (4):

$$\begin{aligned} F_1 &= \{011\} \cap \{010\}, F_2 = \{010\} \cap \{110\}, \\ F_3 &= \{110\} \cap \{100\}, \dots F_5 = \{101\} \cap \{001\}. \end{aligned}$$

After transition along all six arrows of this diagram, trajectories of all points of the face F_0 return to it, each trajectory with its own time. Composition $\Psi: F_0 \rightarrow F_0$ of all these six shifts from face F_m to face F_{m+1} , $m = 0, 1, 2, 3, 4$, and $F_5 \rightarrow F_0$ is called the Poincaré map.

On the face F_0 , let us introduce a coordinate system $(w_1; w_2)$ with the origin at the point $E_3 = (1; 1; 1)$ such that coordinates w_1, w_2 of all points of this face are non-negative: $w_1 = 1 - x_2$; $w_2 = x_3 - 1$. Let the Poincaré map be written by equation

$$\Psi(w_1; w_2) = (\psi_1(w_1; w_2); \psi_2(w_1; w_2)).$$

The main technical result of the papers (Golubyatnikov et al., 2018; Golubyatnikov, Ivanov, 2018) is the following

Lemma 1: a) the Poincaré map is monotonic: if for points $A(v_1; v_2)$ and $B(w_1; w_2)$ relations $v_1 < w_1$ and $v_2 < w_2$, are satisfied then $\psi_1(v_1; v_2) < \psi_1(w_1; w_2)$ and $\psi_2(v_1; v_2) < \psi_2(w_1; w_2)$. For this partial order relation, we use a notation: $A < B$, $\Psi(A) < \Psi(B)$;
b) if w_1 and w_2 are sufficiently small then $w_1 < \psi_1(w_1; w_2)$ and $w_2 < \psi_2(w_1; w_2)$, i. e., $B < \Psi(B)$;
c) at each point of the face F_0 , the first derivatives of the coordinate functions ψ_1 and ψ_2 are strictly positive and their second derivatives are strictly negative.

This implies that the Poincaré map $\Psi: F_0 \rightarrow F_0$ has two fixed points exactly; one of them is the point E_3 which lies at the boundary of F_0 and the other one, denoted by P_* , is contained in the interior of the face F_0 (Golubyatnikov, Ivanov, 2018). Trajectory of the point P_* returns to this point after transition through the blocks of the diagram (4) and, therefore, it is a cycle. Since the map Ψ has just one nontrivial fixed point P_* , the system (3) does not have any other cycles.

In the same paper, for the fixed points E_3 and P_* of the Poincaré map, Jacobian matrices $J_2(E_3)$ and $J_2(P_*)$ were calculated and it was shown that the eigenvalues $\lambda_1(P_*)$, $\lambda_2(P_*)$ of the matrix $J_2(P_*)$ are different, positive and do not exceed one, which means exponential stability of the cycle of the system (3). We denote this cycle discovered in (Golubyatnikov, Ivanov, 2018) by C_3 . Lemma 1 also implies that both these Jacobian matrices are positive, so it is possible to use the Frobenius–Perron theorem (Gantmacher, 1959) in our studies.

Note that the determinant of Jacobian matrix $J_2(E_3)$ is equal to one and for its eigenvalues $\lambda_1(E_3)$, $\lambda_2(E_3)$, relations $\lambda_1(E_3) > 1 > \lambda_2(E_3) > 0$ are true. So, for the map Ψ , hypothesis of Grobman–Hartman theorem (Hartman, 1964) is fulfilled. This implies that in a sufficiently small neighborhood $U(E_3) \subset F_0$ of the point E_3 , the Poincaré map is linearized by some continuous (in general terms, non-smooth) change of variables $(w_1; w_2) \Rightarrow (u_1; u_2)$. In such a coordinate system, $\Psi(u_1; u_2) = (\lambda_1(E_3) \cdot u_1; \lambda_2(E_3) \cdot u_2)$.

For sufficiently small $\varepsilon > 0$, we denote by $T_\varepsilon^2 \subset U(E_3)$ a triangle $0 \leq u_1 + u_2 < \varepsilon$ with one vertex at the point E_3 and let \widehat{F}_0 be a truncated face $F_0 \setminus T_\varepsilon^2$.

Choose two segments $[0, \alpha_1]$ and $[0, \alpha_0] \subset [0, \alpha_1]$ in this neighborhood so that $\alpha_1 = \lambda_1(E_3) \cdot \alpha_0$. Let N_1 and N_0 , respectively, be the right endpoints of these segments, then $\Psi([0, \alpha_0]) = [0, \alpha_1]$ and $\Psi(N_0) = N_1$; in the original coordinate system $(w_1; w_2)$, the segments $[0, \alpha_0]$ and $[0, \alpha_1]$ are represented by arcs $D_0 \subset D_1$ with a common endpoint E_3 . Consider action of iterations of the Poincaré map to these arcs:

$$\Psi(D_0) = D_1 \subset D_2 = \Psi(D_1) \subset D_3 = \Psi(D_2) \subset D_4 \dots$$

The union D_* of infinite sequence of mutually embedded arcs D_k is a continuous monotonic arc connecting the points E_3 and P_* ; after transition along arrows of the diagram (4), trajectories of points of D_* return to this arc: the semi-interval $D_1 \setminus D_0$ passes to semi-interval $D_2 \setminus D_1$ which passes in turn to $D_3 \setminus D_2$, etc. Thus, trajectories of points of the arc D_* generate an invariant (non-smooth) surface Σ^2 bounded by the cycle C_3

in the invariant domain $W_1^3 \subset Q^3$. By the construction, this surface contains the point E_3 .

Starting such constructions of small segments $[N_0, N_1]$ in a neighborhood $U(E_3)$ with points N_0 which do not lie on the axis E_3u_1 and considering the images of these segments under iterations of the Poincaré map Ψ , we obtain a family of continuous monotonic arcs which leave the neighborhood $U(E_3)$ and do not contain the point E_3 . For each pair of points $N_0, N_1 \subset U(E_3) \setminus E_3u_1$ such that $\Psi(N_0) = N_1$, the sequence $N_k = \Psi(N_{k-1})$ tends monotonically to the fixed point P_* of the Poincaré map Ψ (Golubyatnikov et al., 2018). Here, each segment $[N_0, N_1]$ generates, as above, a monotonic arc $D_*(N_0)$ being invariant with respect to the Poincaré map. Trajectories of points of such an arc, in their turn, form an invariant 2D surface $\Sigma^2(N_0)$ which intersects the surface Σ^2 by the cycle C_3 exactly.

In a similar way, one can construct invariant surfaces which do not intersect the neighborhood $U(E_3)$ in the domain W_1^3 . Let $U(P_*) \subset \widehat{F}_0$ be a neighborhood of the nontrivial fixed point P_* , where the map Ψ can be linearized. We save the notations $(u_1; u_2)$ for these linearized coordinates. For sufficiently small $\varepsilon > 0$, the Poincaré map transforms the ellipsis $S_1^1 \subset U(P_*)$ with equation $\lambda_1(P_*)u_1^2 + \lambda_2(P_*)u_2^2 = \varepsilon^2$ to the circle S_0^1 with the equation $u_1^2 + u_2^2 = \varepsilon^2$. Let $I_1(M_0)$ be a segment which joins the point $M_1 \in S_1^1$ with its image $M_0 = \Psi(M_1) \in S_0^1$. All such segments are contained in $U(P_*)$ in the ring between S_0^1 and S_1^1 . Each of these segments generates a sequence of continuous arcs $D_k(M_0)$, they are invariant with respect to the Poincaré map, and $\Psi(D_k(M_0)) = D_{k-1}(M_0)$. For each of these arcs, trajectories of its points generate in W_1^3 an invariant surface bounded by the cycle C_3 .

Theorem 1. *There exists two-dimensional invariant foliation in the invariant domain W_1^3 of the dynamical system (3); its leaves fill W_1^3 and contain the cycle C_3 on their boundaries. One of these leaves contains the point E_3 .*

Four-dimensional dynamical system

Recently, in the papers (Ayupova, Golubyatnikov, 2019; Golubyatnikov, Minushkina, 2021), we considered a four-dimensional block-linear system

$$\frac{dx_1}{dt} = L_1(x_4) - k_1x_1; \quad \frac{dx_r}{dt} = \Gamma_r(x_{r-1}) - k_rx_r; \quad r = 2, 3, 4. \quad (5)$$

In particular case, when $k_j = 1$ for all j , questions of existence, uniqueness, and stability of cycles of such systems were studied in (Glass, Pasternack, 1978). Smooth analogues of similar systems were considered in (Hastings et al., 1977; Mallet-Paret, Smith, 1990).

An invariant domain Q^4 of the system (5) is decomposed by hyperplanes $x_j = 1$ to 16 blocks $\{\varepsilon_1 \varepsilon_2 \varepsilon_3 \varepsilon_4\}$. Blocks of this decomposition listed in the following diagram form an invariant subdomain W_1^4 in the phase portrait of (5)

$$\begin{aligned} & \dots \rightarrow \{1111\} \rightarrow \{0111\} \rightarrow \{0011\} \rightarrow \{0001\} \rightarrow \\ & \{0000\} \rightarrow \{1000\} \rightarrow \{1100\} \rightarrow \{1110\} \rightarrow \{1111\} \rightarrow \dots \end{aligned} \quad (6)$$

The arrows of this diagram show the only possible direction of trajectory transition from one block to another. The subdomain W_1^4 is one of two parts of stratification of the phase

portrait of the system (5). For each block not listed here, trajectories of its points can pass to three adjacent blocks, two of them are contained in W_1^4 , and one is in $Q^4 \setminus W_1^4$ (this is the second part of the stratification mentioned above). Algorithms of construction of such diagrams for the systems of arbitrary dimensions, both smooth and blocks-linear, are described in (Kazantsev, 2015; Kirillova, Minushkina, 2019).

As in previous sections, let us denote by F_0 an intersection of two adjacent blocks $\{1111\} \cap \{0111\}$ in the diagram (6). After eight steps according to its arrows under shifts along trajectories, all points of this three-dimensional face return to F_0 . Let $\Psi_4: F_0 \rightarrow F_0$ be a corresponding Poincaré map, $T_\varepsilon^3 \subset U(E_4)$ be a pyramid $0 \leq u_1 + u_2 + u_3 < \varepsilon$ with the vertex at the point $E_4 = (1; 1; 1; 1)$, and \widehat{F}_0 be a truncated face $F_0 \setminus T_\varepsilon^3$.

In the paper (Golubyatnikov, Minushkina, 2021), it was shown that all statements of Lemma 1 are true for the map Ψ_4 , thus, this map has two fixed points exactly: E_4 and the point Π_* which is contained in the interior of the face F_0 . This means that the invariant domain W_1^4 of the system (5) contains one cycle exactly, let us denote it by C_4 . The following results were also established there.

Lemma 2: a) *the Jacobi matrices $J_3(E_4)$ and $J_3(\Pi_*)$ and their determinants are positive;*
b) $\det J_3(E_4) = \lambda_1(E_4) \cdot \lambda_2(E_4) \cdot \lambda_3(E_4) = 1$;
c) *magnitudes of eigenvalues of the matrix $J_3(\Pi_*)$ are less than one.*

This implies the exponential stability of the cycle C_4 and possibility of linearization of the Poincaré map Ψ_4 in some small neighborhood $U(\Pi_*)$ of its fixed point Π_* . According to the Frobenius–Perron theorem, one of the eigenvalues of the matrix $J_3(\Pi_*)$ is positive and greater than the magnitudes of the remaining eigenvalues. The same applies to the eigenvalues of the matrix $J_3(E_4)$. Let us enumerate the eigenvalues of Jacobi matrices in order of decreasing of their absolute values: $\lambda_1 > |\lambda_2| \geq |\lambda_3|$. Let $(u_1; u_2; u_3)$ be the coordinates where Ψ_4 is linear

$$\Phi(u_1; u_2; u_3) = (\lambda_1(\Pi_*) \cdot u_1; \lambda_2(\Pi_*) \cdot u_2; \lambda_3(\Pi_*) \cdot u_3).$$

As in the case of the system (3), for a sufficiently small $\varepsilon > 0$, the Poincaré map translates the ellipsoid S_1^3 with the equation $\lambda_1(\Pi_*)u_1^2 + |\lambda_2(\Pi_*)|u_2^2 + |\lambda_3(\Pi_*)|u_3^2 = \varepsilon^2$ to the sphere S_0^3 with the equation $u_1^2 + u_2^2 + u_3^2 = \varepsilon^2$.

Theorem 2. *If $a_j > 1$ for all $j = 1, 2, 3, 4$, and the Jacobi matrix $J_3(E_4)$ of the Poincaré map does not have eigenvalues with unit module then there exists an invariant foliation in the domain W_1^4 ; its leaves fill this invariant domain and contain the cycle C_4 . One of these leaves contains the point E_4 .*

Dynamical systems of higher dimensions

In the papers (Gaidov, Golubyatnikov, 2014; Ayupova, Golubyatnikov, 2021), we considered a five-dimensional block linear dynamical system

$$\dot{x}_1 = L_1(x_5) - k_1x_1; \quad \dot{x}_2 = L_2(x_1) - k_2x_2; \quad \dots \quad \dot{x}_5 = L_5(x_4) - k_5x_5, \quad (7)$$

for which, as in previous sections, an invariant domain $Q^5 = [0, a_1] \times [0, a_2] \times \dots \times [0, a_5]$ and its decomposition to blocks

by the hyperplanes $x_j = 1$ were constructed. Ten blocks of this decomposition form a stratum $W_1^5 \subset Q^5$ which is invariant with respect to shifts along trajectories of the system (7) passing through the blocks according to arrows of a cyclic diagram, similar to (4) and (6):

$$\begin{aligned} \dots \rightarrow \{10101\} \rightarrow \{00101\} \rightarrow \{01101\} \rightarrow \{01001\} \rightarrow \\ \{01011\} \rightarrow \{01010\} \rightarrow \{11010\} \rightarrow \{10010\} \rightarrow \\ \{10110\} \rightarrow \{10100\} \rightarrow \{10101\} \rightarrow \dots \end{aligned}$$

Points of the four-dimensional face $F_0^4 = \{10101\} \cap \{00101\}$ under shifts along their trajectories after ten steps along the arrows of the diagram return to the face F_0^4 .

For such a Poincaré map $\Psi_5: F_0^4 \rightarrow F_0^4$, an analogue of Lemma 1 implies that the face F_0^4 contains two fixed points of this map exactly: the point $E_5 = (1; 1; 1; 1; 1)$ and a point Π_*^5 in the interior of this face. The domain W_1^5 contains one cycle exactly. Let us denote it by C_5 . This cycle is stable and passes through the point Π_*^5 (Ayupova, Golubyatnikov, 2021).

As in previous sections, an analogue of Lemma 2 holds: Jacobi matrices $J_4(E_5)$, $J_4(\Pi_*^5)$ and their determinants are positive, $\det J_4(E_5) = 1$.

The magnitudes of eigenvalues of the matrix $J_4(\Pi_*^5)$ do not exceed one. In the case when these Jacobi matrices do not have any eigenvalues modulo equal to 1, construction of the invariant surface $\Sigma^2 \subset W_1^5$ with the boundary C_5 and an invariant foliation in the domain W_1^5 is carried out exactly in the same way as above.

In the paper (Golubyatnikov, Gradov, 2021), conditions under which a non-invariant stratum $Q^5 \setminus (W_1^5 \cup \{00000\} \cup \{11111\})$ of the phase portrait of the five-dimensional system (7) contains one more of its cycle were established.

Similar constructions can be done for a block-linear analogue of the six-dimensional Elowitz–Leibler system (Elowitz, Leibler, 2000) studied in (Minushkina, 2021; Golubyatnikov, Minushkina, 2022)

$$\begin{aligned} \dot{m}_1 = L_1(p_3) - k_1 m_1; \quad \dot{p}_1 = \Gamma_1(m_1) - l_1 p_1; \quad \dot{m}_2 = L_2(p_1) - k_2 m_2; \\ \dot{p}_2 = \Gamma_2(m_2) - l_2 p_2; \quad \dot{m}_3 = L_3(p_2) - k_3 m_3; \quad \dot{p}_3 = \Gamma_3(m_3) - l_3 p_3. \end{aligned} \quad (8)$$

Here, the variables m_j and p_j denote concentrations of three mRNAs and proteins TetR, LacI and λ cl, corresponding to them (Elowitz, Leibler, 2000; Kolesov et al., 2016).

An invariant domain $Q^6 = \Pi_{j=1}^3 [0, a_j] \times [0, b_j]$, where b_j are the maximum values of step functions Γ_j divided by the coefficients l_j , $j = 1, 2, 3$, is decomposed by six hyperplanes $m_j = 1$, $p_j = 1$, $j = 1, 2, 3$, to 64 blocks which form a stratification of Q^6 to three subdomains, W_1^6 , W_3^6 , and W_5^6 , with different qualitative trajectory behavior.

The domain W_5^6 consists of 12 blocks, from which trajectories can transit to 5 adjacent blocks. In the symmetric case when $k_j = l_j = 1$, there are no cycles in this subdomain. However, W_5^6 contains a two-dimensional invariant surface consisting of piecewise linear trajectories attracting by the point $E_6 = (1; 1; 1; 1; 1; 1)$ in a spiral way.

In the domain W_1^6 formed by 12 blocks, from which trajectories can enter one adjacent block only, the Poincaré map contains a unique non-trivial fixed point Π_*^6 , the trajectory of

which is a stable limit cycle for all trajectories in this domain (Golubyatnikov, Minushkina, 2022).

In the domain W_3^6 which consists of 40 blocks, state transition diagram has a more complicated combinatorial structure. At present time, transitions of trajectories from one block to another in this subdomain have not been studied completely yet.

For smooth analogues of the dynamical system (8), the uniqueness of equilibrium point was established in (Ayupova et al., 2017). As in the case of block linear systems, hyperplanes passing through the equilibrium point and being parallel to coordinate ones decompose the invariant domain Q^6 to 64 blocks. If a linearization matrix of such smooth system in its equilibrium point has eigenvalues with positive and negative real parts and does not have any imaginary eigenvalues then the invariant domain W_1^6 contains a cycle C_6 of this system (Ayupova et al., 2017). In the paper (Kirillova, 2020), conditions of existence of an invariant surface bounded by the cycle C_6 in the domain W_1^6 were obtained.

Results of numerical experiments

The lefthand part of Figure shows 100 trajectories of the dynamical system (3). Each of these trajectories is contained in a corresponding leaf of the foliation in W_1^3 near the invariant surface Σ^2 . The values of parameters of this system are: $k_1 = 0.4$; $k_2 = 0.3$; $k_3 = 0.6$; $a_1 = 1.3$; $a_2 = 1.4$; $a_3 = 1.7$. The initial data are chosen in a random way in a rectangular neighborhood of the point E_3 . The righthand part of this Figure shows results of similar experiments with a smooth analogue of the system (3):

$$\frac{dx}{dt} = \frac{10}{1+z^3} - x; \quad \frac{dy}{dt} = \frac{10}{1+x^3} - y; \quad \frac{dz}{dt} = \frac{10}{1+y^3} - z.$$

Here, one can clearly see its invariant surface.

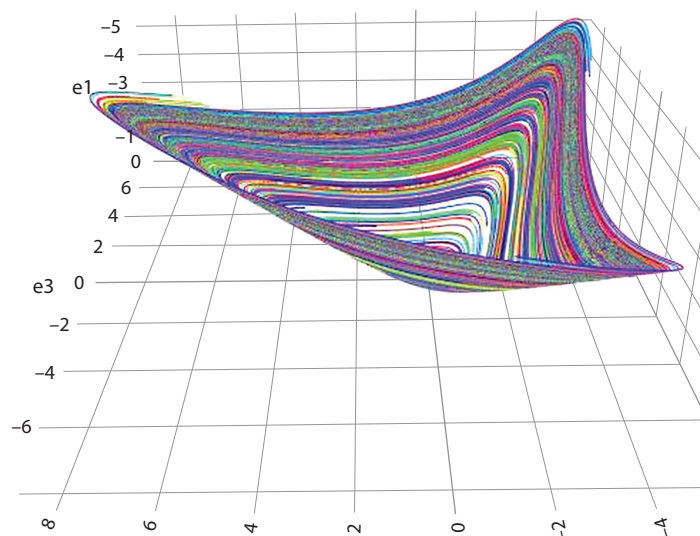
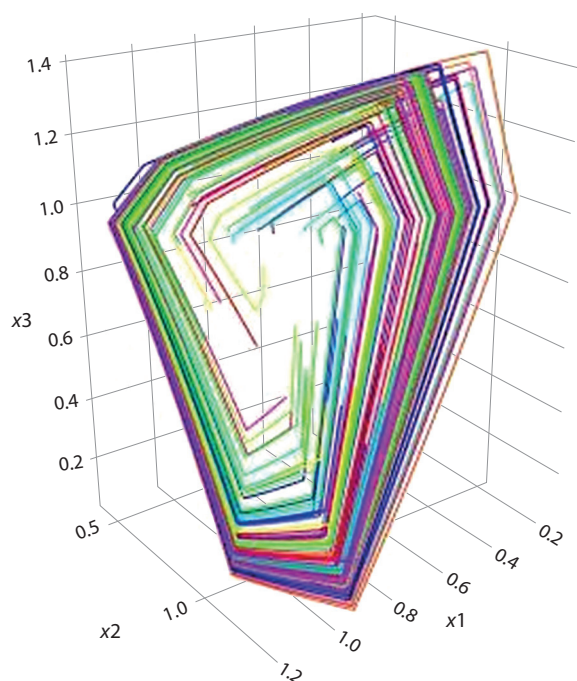
It was shown in (Golubyatnikov et al., 2018; Ayupova, Golubyatnikov, 2021; Golubyatnikov, Minushkina, 2021; Minushkina, 2021) that trajectories of block-linear dynamical systems (3), (5), (7), (8) are piecewise smooth, the discontinuities of their derivatives are located on the planes $x_j = 1$, this is clearly seen on the left part of Figure.

In order to perform numerical simulations of trajectories of (3), we have developed a software project using the R programming language (<https://www.r-project.org/>) and the Shiny package (<https://shiny.rstudio.com/>). The source code is available on GitHub: <https://github.com/AndreyAkinshin/pwLLL>.

The simulations are performed in the cloud; the results are described at <https://aakinshin.net/posts/dscs2/>. The library ggplot (<https://ggplot2.tidyverse.org/>) is used here, as well as the package deSolve (<http://desolve.r-forge.r-project.org/>) that contains integration routines previously used to simulate other systems of gene networks. The user interface allows one to specify all parameters of the system (3).

Conclusion

In this paper, we have described a construction of invariant foliations, i. e. the families of invariant two-dimensional surfaces in phase portraits of low-dimensional block-linear models of circular gene networks. It was shown that on each leaf of



Results of numerical experiments with trajectories of the 3D systems.

these foliations, trajectories of all its points are repelled by the boundary of the central part of the phase portrait and they are attracted by the limit cycle which describes an oscillating functioning of the corresponding gene network. Theorem 1 is illustrated by numerical experiments.

For the kinetic dynamical systems under consideration, the leaves of invariant foliations in the phase portraits play the role of level surfaces of collections of motion integrals studied in classical mechanics (Poincaré, 1892; Arnold, 1989). Reduction of dimensions of invariant subsets in the phase portraits allows us to give a digestible description of trajectories behavior and, in particular, simplifies considerably numerical experiments with such gene networks models (Likhoshvai et al., 2020). Construction of the foliations mentioned above and investigation of their geometric properties can be useful in studies of dynamical characteristics of more complicated models of gene networks functioning when a description of a big system is given on the basis of known results on its subsystems which have a simpler structure.

References

- Akinshin A.A., Golubyatnikov V.P., Golubyatnikov I.V. On some multidimensional models of gene network functioning. *J. Appl. Ind. Math.* 2013;7(3):296-301. DOI 10.1134/S1990478913030022.
- Arnold V.I. *Mathematical Methods of Classical Mechanics*. 2-d ed. Springer, 1989.
- Ayupova N.B., Golubyatnikov V.P. On the uniqueness of a cycle in an asymmetric three-dimensional model of a molecular repressilator. *J. Appl. Ind. Math.* 2014;8(2):153-157. DOI 10.1134/S199047891402001X.
- Ayupova N.B., Golubyatnikov V.P. Structure of phase portrait of a piecewise-linear dynamical system. *J. Appl. Ind. Math.* 2019;13(4):606-611. DOI 10.1134/S1990478919040033.
- Ayupova N.B., Golubyatnikov V.P. On a cycle in a 5-dimensional circular gene network model. *J. Appl. Ind. Math.* 2021;15(3):376-383. DOI 10.1134/S1990478921030029.
- Ayupova N.B., Golubyatnikov V.P., Kazantsev M.V. On the existence of a cycle in an asymmetric model of a molecular repressilator. *Num. Anal. Appl.* 2017;10(2):101-107. DOI 10.1134/S199542391702001X.
- Elowitz M.B., Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature*. 2000;403:335-338. DOI 10.1038/35002125.
- Gaidov Yu.A., Golubyatnikov V.P. On cycles and other geometric phenomena in phase portraits of some nonlinear dynamical systems. In: *Geometry and Applications. Springer Proc. Math. Stat.* 2014;72:225-233. DOI 10.1007/978-3-319-04675-4_10.
- Gantmacher F.R. *Applications of the Theory of Matrices*. New York; London: Interscience Publ., 1959.
- Glass L., Pasternack J.S. Stable oscillations in mathematical models of biological control systems. *J. Math. Biol.* 1978;6:207-223.
- Glyzin S.D., Kolesov A.Yu., Rozov N.Kh. Buffering in cyclic gene networks. *Theor. Math. Phys.* 2016;187(3):935-951. DOI 10.1134/S0040577916060106.
- Golubyatnikov V.P., Gradov V.S. Non-uniqueness of cycles in piecewise-linear models of circular gene networks. *Sib. Adv. Math.* 2021;31(1):1-12. DOI 10.3103/S1055134421010016.
- Golubyatnikov V.P., Ivanov V.V. Uniqueness and stability of a cycle in three-dimensional block-linear circular gene network models. *Sibirskii Zhurnal Chistoi i Prikladnoi Matematiki = Siberian Journal of Pure and Applied Mathematics*. 2018;18(4):19-28. DOI 10.33048/PAM.2018.18.402. (in Russian)
- Golubyatnikov V.P., Ivanov V.V., Minushkina L.S. On existence of a cycle in one asymmetric gene network model. *Sibirskii Zhurnal Chistoi i Prikladnoi Matematiki = Siberian Journal of Pure and Applied Mathematics*. 2018;18(3):27-35. DOI 10.17377/PAM.2018.18.4. (in Russian)
- Golubyatnikov V.P., Minushkina L.S. Monotonicity of the Poincaré mapping in some models of circular gene networks. *J. Appl. Ind. Math.* 2019;13(3):472-479. DOI 10.1134/S1990478919030086.
- Golubyatnikov V.P., Minushkina L.S. Combinatorics and geometry of circular gene networks models. *Pisma v Vavilovskii Zhurnal Genetiki i Selektii = Letters to Vavilov Journal of Genetics and Breeding*. 2020;6(4):188-192. DOI 10.18699/Letters2020-6-24.

- Golubyatnikov V.P., Minushkina L.S. On uniqueness and stability of a cycle on one gene network. *Siberian Electronic Mathematical Reports*. 2021;18(1):464-473. DOI 10.33048/semi.2021.18.032.
- Golubyatnikov V.P., Minushkina L.S. On uniqueness of a cycle in one circular gene network model. *Sib. Math. J.* 2022;63(1):79-86. DOI 10.1134/S0037446622010062.
- Hartman Ph. Ordinary Differential Equations. New York: John Wiley, 1964.
- Hastings S., Tyson J., Webster D. Existence of periodic solutions for negative feedback cellular control systems. *J. Diff. Eqn.* 1977;25: 39-64.
- Ivanov V.V. Attracting limit cycle of an odd-dimensional circular gene network model. *J. Appl. Ind. Math.* 2022;16(3):409-415. DOI 10.1134/S199047892203005X.
- Kazantsev M.V. On some properties of the domain graphs of dynamical systems, *Sibirskii Zhurnal Industrialnoi Matematiki = Siberian Journal of Applied and Industrial Mathematics*. 2015;18(4):42-48. DOI 10.17377/sibjim.2015.18.405. (in Russian)
- Kirillova N.E. On invariant surfaces in gene network models. *J. Appl. Ind. Math.* 2020;14(4):666-671. DOI 10.1134/S1990478920040055.
- Kirillova N.E., Minushkina L.S. On the discretization of phase portraits of dynamical systems. *Izvestiya Altayskogo Gosudarstvennogo Universiteta = Izvestiya of Altai State University*. 2019;108(4):82-85. DOI 10.14258/izvasu(2019)4-12. (in Russian)
- Kolesov A.Yu., Rozov N.Kh., Sadovnichii V.A. Periodic solutions of travelling-wave type in circular gene networks. *Izvestiya: Mathematics*. 2016;80(3):523-548.
- Likhoshvai V.A., Golubyatnikov V.P., Khlebodarova T.M. Limit cycles in models of circular gene networks regulated by negative feedback loops. *BMC Bioinformatics*. 2020;21(Suppl. 11):255. DOI 10.1186/s12859-020-03598-z.
- Mallet-Paret J., Smith H. The Poincaré–Bendixson theorem for monotone cyclic feedback systems. *J. Dynam. Diff. Eqns.* 1990;2(4): 367-421.
- Minushkina L.S. Phase portraits of a block-linear dynamical system in a model for a circular gene network. *Matematicheskiye Zametki SVFU = Mathematical Notes of NEFU*. 2021;28(2):34-46. DOI 10.25587/SVFU.2021.60.20.003. (in Russian)
- Poincaré H. Les Méthodes Nouvelles de la Mécanique Céleste. T. I. Solutions Périodiques. Non-existence des Intégrales Uniformes. Solutions Asymptotiques. Paris: Gauthier-Villars et fils, 1892.

ORCID ID

V.P. Golubyatnikov orcid.org/0000-0002-9758-3833

Acknowledgements. The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project No. FWNF-2022-0009 “Inverse problems of natural science and tomography problems”).

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

Received August 11, 2022. Revised September 20, 2022. Accepted September 21, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Validation of a face image assessment technology to study the dynamics of human functional states in the EEG resting-state paradigm

A.N. Savostyanov^{1, 2, 3} , E.G. Vergunov², A.E. Saprygin^{1, 2}, D.A. Lebedkin^{2, 3}

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia

³ Institute for the Humanities of Novosibirsk State University, Novosibirsk, Russia

 a-sav@mail.ru

Abstract. The article presents the results of a study aimed at finding covariates to account for the activity of implicit cognitive processes in conditions of functional rest of the subjects and during them being presented their own or someone else's face in a joint analysis of EEG experiment data. The proposed approach is based on the analysis of the dynamics of the facial muscles of the subject recorded on video. The pilot study involved 18 healthy volunteers. In the experiment, the subjects were sitting in front of a computer screen and performed the following task: sequentially closed their eyes (three trials of 2 minutes each) and opened them (three trials of the same duration between periods of closed eyes) when the screen was either empty or when it was showing a video recording of their own face or the face of an unfamiliar person of the same gender as the participant. EEG, ECG and a video of the face were recorded for all subjects. In the work a separate subtask of the study was also addressed: validating a technique for assessing the dynamics of the subjects' facial muscle activity using the recorded videos of the "eyes open" trials to obtain covariates that can be included in subsequent processing along with EEG correlates in neurocognitive experiments with a paradigm that does not involve the performance of active cognitive tasks ("resting-state conditions"). It was shown that the subject's gender, stimulus type (screen empty or showing own/other face), trial number are accompanied by differences in facial activity and can be used as study-specific covariates. It was concluded that the analysis of the dynamics of facial activity based on video recording of "eyes open" trials can be used as an additional method in neurocognitive research to study implicit cognitive processes associated with the perception of oneself and other, in the functional rest paradigm.

Key words: neurocognitive studies; own and other face; EEG correlates; covariates; implicit cognitive processes; self-perception.

For citation: Savostyanov A.N., Vergunov E.G., Saprygin A.E., Lebedkin D.A. Validation of a face image assessment technology to study the dynamics of human functional states in the EEG resting-state paradigm. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):765-772. DOI 10.18699/VJGB-22-92


Апробация технологии оценки мимики лиц для изучения динамики функциональных состояний человека в ЭЭГ-парадигме покоя

А.Н. Савостьянов^{1, 2, 3} , Е.Г. Вергунов², А.Е. Сапрыгин^{1, 2}, Д.А. Лебедин^{2, 3}

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия

³ Новосибирский национальный исследовательский государственный университет, Гуманитарный институт, Новосибирск, Россия

 a-sav@mail.ru

Аннотация. В статье представлены результаты исследования, направленного на поиск ковариат для учета деятельности имплицитных когнитивных процессов в условиях функционального покоя испытуемых и при демонстрации им собственного или чужого лица в совместном анализе данных ЭЭГ-эксперимента. Предлагаемый подход основан на анализе динамики мышц лица испытуемого по видео. В пилотном исследовании приняли участие 18 здоровых добровольцев. В эксперименте испытуемые, сидя перед экраном, последовательно закрывали глаза (три пробы по 2 минуты) и открывали их (также три пробы между периодами закрытых глаз) либо перед пустым экраном, либо перед экраном с демонстрацией видеозаписи их собственного лица или лица незнакомого им человека такого же пола, что и участник. У всех испытуемых регистрировали ЭЭГ, ЭКГ и вели запись видео лица. В работе решали отдельную подзадачу эксперимента: апробацию методики оценки динамики

активности мышц лица испытуемых по их видео с открытыми глазами для получения ковариат, которые можно включать в последующую обработку совместно с ЭЭГ-коррелятами в нейрокогнитивных экспериментах с парадигмой, не предполагающей выполнение активных когнитивных заданий (resting-state conditions). Показано, что пол испытуемого, статус экрана (пустой, собственное/чужое лицо), номер пробы связаны с различиями в мимической активности лица и могут выступать искомыми ковариатами. Сделан вывод, что анализ динамики мимической активности по видео с открытыми глазами может быть дополнительным методом в нейрокогнитивных исследованиях для изучения имплицитных когнитивных процессов, связанных с восприятием изображения себя и другого, в парадигме функционального покоя.

Ключевые слова: нейрокогнитивные исследования; свое и чужое лицо; ЭЭГ-корреляты; ковариаты; имплицитные когнитивные процессы; самовосприятие.

Introduction

Technologies of neurocognitive studies are most often based on the use of various approaches to recording the brain activity of experiment participants using techniques such as EEG or fMRI (Bringas-Vega et al., 2022). In the last two decades (Biswal, 2012; Snyder, Raichle, 2012) the researchers' interest has been focused on the functional states of the brain observed in the absence of exogenous cognitive or emotional load, that is, in the experimental paradigm of "resting-state conditions".

In a series of studies, it was shown that the functional states of the brain at rest reflect the individual characteristics of the subjects, including their gender (Volf et al., 2015), age (Privodnova et al., 2020; Engemann et al., 2022), genetic features (Proshina et al., 2018), sociocultural affiliation (Knyazev et al., 2012), climatic and geographical living conditions (Milakhina et al., 2020), psychological personality traits (Kabbara et al., 2020) and predisposition to affective disorders (Greicius et al., 2007). However, the problem of using neuroimaging techniques consists in the high variability of resting-state brain activity characteristics in healthy subjects (Li et al., 2022). A comparative study by M. Li and colleagues, performed on a sample of more than 1500 participants in nine countries, showed that the resting-state EEG characteristics of a healthy person vary greatly depending both on the characteristics of the subjects and on the conditions in which imaging sessions take place which are not specified in the experimental paradigm (Li et al., 2022). At the same time, formally the same EEG recording conditions (closed eyes without external mental load) can give different results depending on the part of the world and the period of the year the EEG was recorded in.

One of the factors that significantly changes the functional states of the brain at rest is the presence or absence of the person's thoughts about themselves during the period of registration of their brain activity. In the work (Knyazev et al., 2012) it was shown that thinking about oneself induces increased activity of the default mode brain network. At the same time, the functional organization of the default system under these conditions demonstrated significant intercultural differences when comparing subjects from Novosibirsk and Taiwan.

In the case of fMRI, an additional factor is the person's response to the very situation of placing them in the scanner. The fMRI recording is done while the person is lying in a confined tube with sound-induced noise and limited mobility, and sometimes contrast agent injection is required. Obviously, some people react to such conditions as a stressor, while other people perceive these conditions differently, which causes a

wide spread in the assessments' results of the subjects' functional state. Hence, the task arises: on the basis of additional methods, to find such correlates (or covariates) that, during subsequent analysis, together with the results of EEG or fMRI examinations, will allow to more precisely account for the psychophysiological state of the subject.

In the case of an experimental paradigm using stimuli to induce the desired state of the participants, the assessment of such a state is done by analyzing behavioral indicators (for example, the accuracy/speed of response to external stimuli), but in the case of the resting-state studies, this is not possible.

Another method consists in the usage of psychological questionnaires that the participant is asked to complete before or after the experiment session. Questionnaire indicators are used as variables to assess the subjective states of a person under experimental conditions or their personality traits. However, this method is limited by the sincerity of the test subject and their ability for adequate self-assessment, which can be pronounced in the case of neuropsychiatric diseases.

In our pilot study, we propose an approach using covariates that can be obtained from the dynamics of facial muscle activity recorded on video and are associated with the psychophysiological state of the participants in the EEG experiment. The analysis of facial muscle activity in psychophysiology has been tested (Nikolaeva, Vergunov, 2021), but has not been used for joint analysis with EEG data.

We test the hypothesis that the subjects' facial activity dynamics and the duration of the eyes screen fixation in resting-state activity sessions with the absence of explicit experimental tasks differ depending on the factors such as the subject's gender, the demonstration of a blank screen or a screen with a video of their own face or a face of another person of the same gender, the order of experiment stages ("blank screen", "own face", "other face").

The participants were subjected to complex psychological testing to assess their personality traits with co-registration of EEG, ECG and video recording of facial activity. However, within the framework of this study, we will not present the results of EEG, ECG, and psychometry, leaving them for future joint analysis with the identified covariates at subsequent stages of the experiment.

Materials and methods

Sample description. The experiments involved 18 volunteers (8 men and 10 women, mean age 19.5 ± 1.3 years), all students of Novosibirsk State University. Before the survey,

all participants signed an informed consent form. In addition, all subjects completed a questionnaire for the presence of psychiatric or neurological diseases, a questionnaire for well-being before the examination, and for the use of alcohol or psychoactive substances. The exclusion criteria were:

- certain established medical diagnoses;
- use of drugs or psychotropic medications;
- a state of alcoholic intoxication or severe psychological stress;
- violation of the instructions during the experiment session (covering part of the face with a hand, sudden movements, changing the posture so that part of the face goes out of the camera frame, etc.).

Experiment session procedure. During the experiment session, the subjects sat in a chair in a soundproof chamber with subdued lighting. An EEG helmet was placed on the participant's head, and electrodes were attached to the left arm and both legs for ECG recording. The subjects were informed that in the process of recording EEG and ECG, a video recording of their face was being made. The protocol of the experiment was approved by the ethical committee of the Scientific Research Institute of Neurosciences and Medicine in accordance with the ethical standards of the Declaration of Helsinki for biomedical research.

Participants were instructed to minimize movement of their arms, legs, and head. During the EEG recording session they had to, on command given by the computer, open or close their eyes. Participants were not specifically required to focus their eyes on the screen, but they were not prohibited from doing so. Each participant was tested in three different conditions:

- a) background recording with alternating eye closing/opening (3 trials of each type for 2 minutes), in which there were no images on the computer screen;
- b) recording with opening and closing of the eyes, in which a video recording of the participant's own face, made earlier during condition (a), was shown on the screen (3 trials of each type for 2 minutes);
- c) recording with opening and closing of the eyes, in which the participant was shown a video recording of the face of a person he did not know, but of the same gender as the participant (3 trials of each type for 2 minutes).

All participants were examined in all three conditions. The first condition has always been the (a) condition, i.e., recording without additional external stimulation, for half of the participants the second condition was (b) (own face), and the third was (c) (another face), and for the other half of the participants, on the contrary, the second was the condition (c), and the third was (b).

In between these recordings, participants performed active experimental tasks – solving linguistic tests for finding syntactic errors in sentences between the first and second examinations (approximately 25 minutes) and performing motor tests in the stop-signal paradigm (approximately 12 minutes) between the second and third examinations.

Before the first experimental condition, all participants filled out the Russian version of the C. Spielberger questionnaire to assess the level of situational anxiety (Khanin, 1976).

After completion of the first condition, the C. Spielberger questionnaire was filled out again to assess whether participation in the survey affects the level of situational anxiety. In addition, after completing each of the experimental conditions, the participants filled out a G.G. Knyazev questionnaire on well-being during the EEG recording (Knyazev et al., 2012). Thus, each participant filled out the C. Spielberger questionnaire twice (before and after the first experimental condition), and G.G. Knyazev three times (after each experimental condition).

Our proposed study design allows to control the factors that may accompany implicit cognitive processes taking place during presentation of faces (one's own and other's) or a blank screen:

- features of the motor units activity for the face muscles (AU) according to Facial Action Coding System (FACS);
- features of time distribution in the test in relation to the subject's gaze fixation on the screen;
- features of perception for subjects of different genders;
- features of perception for the first and subsequent conditions;
- individual specificity of implicit cognitive processes associated with the personality traits of the subjects, such as the level of anxiety.

Note that the analysis of the last factor (individual specificity) is not included in the objectives of this study. Later, a joint analysis of the results of psychological questionnaires with the results of clustering statistics for this factor will be used for psychophysiological profiling of the subjects.

Method for assessing the expression of facial muscles.

Specialized software tools, including those in open access, are being widely developed to assess the expression of facial muscles of the subject from video. The OpenFace framework was used in this study – an open access solution that allows to highlight a person's face from an image, from a sequence of images or from a video stream (Saprygin et al., 2022). A video stream was recorded on a regular computer video camera (webcam) when the subjects performed tasks, then, based on the regression model, facial motor units (AU) were identified using the FACS system (facial action coding system) and the dynamics of their activity during tests with open eyes were analyzed (Fig. 1). Empirically, it was found that the dynamics of AU during a period of about two minutes of immobile sitting of the subject is best characterized not by the average value or standard deviation (a large number of small random changes create “noise”), but by the range of values. Therefore, exactly the range of expression values for each AU was included into the analysis.

The OpenFace framework is based on the CLM (constrained local model) approach. The pilot software developed by the authors based on OpenFace_GUI allows real-time visualization of a set of features provided by 3D models of the OpenFace framework (coordinates of key points of the face, position and angles of the head in space, direction of gaze). The OpenFace framework consists of three main parts: 1) C++ code in which the main analytical flow is implemented; 2) files of pre-trained models for face detection, detection and

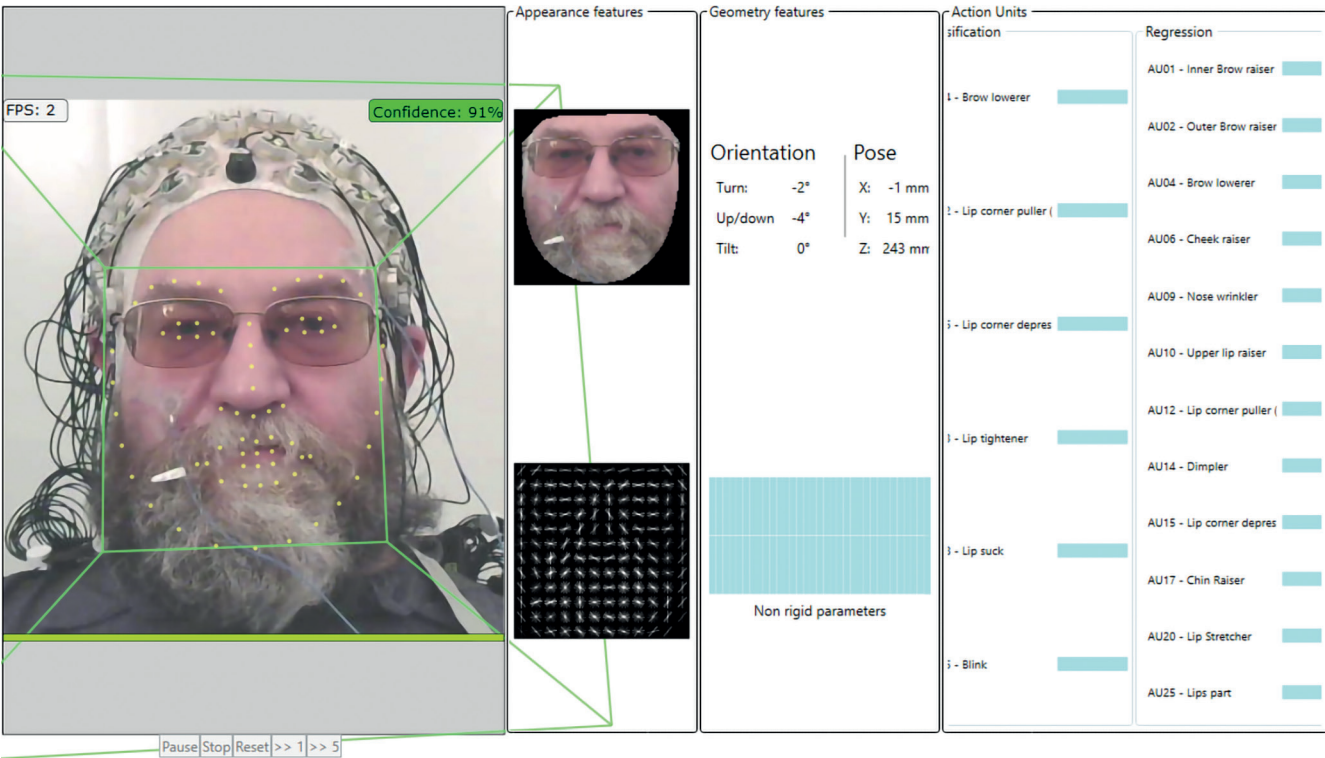


Fig. 1. A screenshot of the program for the processing of facial expressions of the participant’s face from the video with added complex analysis-wise elements (dark glasses, beard, mustache, cap with electrodes for EEG recording).

FPS – processing speed, number of frames per second; confidence – the level of reliability (green indicates an acceptable level); appearance features – features of facial expressions recognized by the program after bringing the face to a vertical position; geometry features – 3D geometry of the position and orientation of the face; action units is the activity of motor units for facial muscles (AU according to FACS); classification – AU values obtained by the classification method (not used in this study); regression – AU values obtained by the regression method (see the Table); orientation – angular 3D indicators of face orientation; turn – face rotation (left+, right–); up/down – tilt of the face (up+, down–); tilt – tilt of the face to the shoulder (left+, right–); pose – linear indicators of the position of the face; X, Y, Z – coordinates of the center of the face (in mm); non rigid parameters – soft parameters; pause, stop, reset, >> 1, >> 5 – frame/video player control buttons.

tracking of key points of the face, calculation of motor units;
3) Matlab code to create your own model files.

Model files are created using a wide variety of training datasets. The OpenFace framework code is open source and available under the GNU license: <https://github.com/Tadas-Baltrusaitis/OpenFace>.

Mathematical foundations of the model. PLS-analysis is a method of obtaining projections on latent structures, the original name of which is “partial least squares method”. An effective tool for PLS analysis is 2B-PLS models (2B-PLS, two-block PLS) (Rohlf, Corti, 2000). 2B-PLS models being applied to the study of implicit cognitive processes reveal deep independent (orthogonal) “latent structures” (psycho-physiological mechanisms) simultaneously for two different blocks (matrices B1 and B2) of multidimensional indicators (Kovaleva et al., 2019).

When constructing 2B-PLS models, the data series are centered, both blocks are scaled and rotated to obtain the maximum covariance between the score matrices (B1- and B2-score), which are projections of the matrices B1 and B2 onto the desired latent structures. This is the main difference between 2B-PLS and PCA (principal component analysis, the method of principal components), which allows you to build models only of a “single-component” type. For example,

one block can contain feature variables (consisting only of “0” and “1”, the variance is minimal), and the other-rows of instrumental data (in which the variance is much larger than that of the features).

The latent structures obtained in the 2B-PLS model are described using orthogonal load matrices (B1- and B2-loadings). Rows in matrices B1 and B2 are objects’ data, columns are the indicators. Thus, indicators act as initial coordinate axes (including those correlated with each other), and can be considered as “explicit structures”, each of which determines a certain (usually small) amount of total variance. The purpose of the 2B-PLS model is to find a system of pairs of axes for both blocks at once, which express the maximum covariance pattern (Polunin et al., 2019). At the same time, the load matrices are the transition matrices from the original “explicit structures” to the newfound “latent structures”.

As a result of applying a 2B-PLS model, we get the number of latent structures (new coordinate axes), which is equal to the minimum number of variables from the two blocks of initial data. Note that the ratios for raw data structures in blocks remain the same after any number (and order) of application of operations such as centering, scaling and rotation, which are applied in PLS models or PCA models. Thus, the structure of the raw data is completely preserved,

Blocks of variables for the 2B-PLS model

Variable	Block
AU01 – frontalis pars medialis, inner brow raiser	No. 1
AU02 – frontalis pars lateralis, outer brow raiser	No. 1
AU04 – depressor glabellae, brow lowerer	No. 1
AU06 – orbicularis oculi pars orbitalis, cheek raiser	No. 1
AU09 – levator labii superioris alaeque nasi, nose wrinkler	No. 1
AU10 – levator labii superioris, upper lip raiser	No. 1
AU12 – zygomaticus major, lip corner puller	No. 1
AU14 – buccinator, dimpler	No. 1
AU15 – triangularis, lip corner depressor	No. 1
AU17 – mentalis, chin raiser	No. 1
AU20 – risorius with platysma, lip stretcher	No. 1
AU25 – depressor labii inferioris, lips part	No. 1
t – the time proportion (from the whole duration of the presentation) of the subjects' eyes fixation on the screen upon presentation of stimuli	No. 1
n1–n3 – indicator for stimulus presentation order number	No. 2
f – indicator for female sex	No. 2
m – indicator for male sex	No. 2
fn – stimulus indicator (screen without a face)	No. 2
tf – stimulus indicator (screen with other face of the same gender)	No. 2
wf – stimulus indicator (screen with own face)	No. 2
s6–s49 – indicators of individual specificity (subject codes)	No. 2

Note. AU classification is according to Facial Action Coding System (FACS).

while the tools of the least squares method (ordinary least squares, OLS) in some cases can lead to alteration of the original structure.

As a result of building a 2B-PLS PLS model, all information from the initial data series (the number of which can be hundreds or more) is collected into the first few independent latent structures. 2B-PLS model allows for a situation where the number of variables is greater than the number of objects, as well as for the cross-correlation of the initial data. Moreover, the data series can be linear combinations of each other (Ränner et al., 1994).

Results and discussion

A 2B-PLS model was built, the blocks of which included the following variables, which are series of instrumental data (13 variables, block No. 1) and series of features (26 variables, block No. 2) (see the Table). Accordingly, 13 latent structures were obtained.

As follows from the “scree” plot for latent structures of the constructed 2B-PLS model (Fig. 2), the first inflection of the graph falls on structure No. 2. Thus, structure No. 1 (before

the first inflection) will reflect the general features of implicit cognitive processes (as it is confirmed by the proportion of observed total variance caused by it).

The second inflection of the graph falls on structure No. 4. Thus, for structures No. 2 and 3, the particular specificity of implicit cognitive processes will be defining. In the subsequent structures, the noise component grows simultaneously with a decrease in the share of the described total variance, however, we will also consider structure No. 4 – it causes more than 5 % of the total variance.

Later, the analysis of the results of psychological questionnaires, together with the results of clustering for the structures we obtained, can be used for the purposes of psychophysiological profiling of the subjects. Hence the conclusion that for subsequent profiling in the EEG experiment, it is necessary to assess the influence of individual differences of the subjects in their implicit cognitive processes when being presented with their own or someone else's face.

The first four latent structures describe 85.4 % of the total variance and the defining features are gender, stimulus type, and trial order.

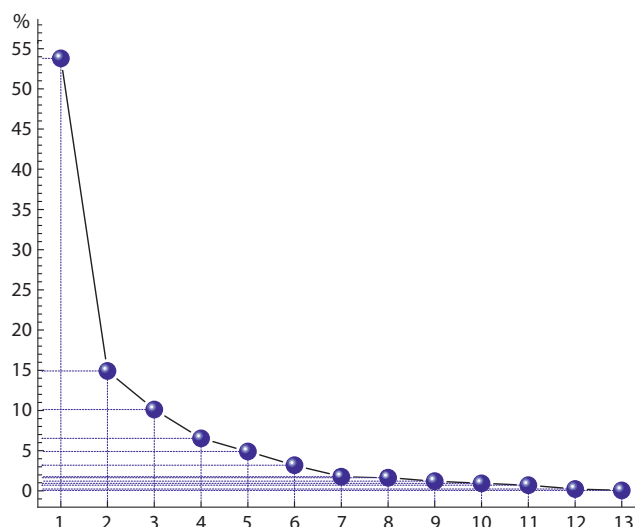


Fig. 2. Scree plot for latent structures of the constructed 2B-PLS model. X-axis is the numbers of latent structures; Y-axis is the share of the observed total variance described by them.

According to Fig. 3, the first structure describes 53.8 % of the total variance and is determined by the proportion of the time the subject's gaze is fixed on the screen, the activity of the buccinator and risorius muscles, gender characteristics, and the perception of all first trials. Hence, the perception of all the first samples is accompanied by an increase in activity of the buccinator and risorius muscles and a decrease in the proportion of the time of the gaze fixation on the screen in girls, and in boys – by a decrease in activity of said muscles and an increase in the proportion of the time of screen-fixed gaze.

The second structure describes 14.9 % of the total variance and is determined by the proportion of time the subject's gaze is fixed on the screen, the activity of the cheek raiser muscle, and signs of the type of stimuli (see Fig. 3). Hence, the perception by all subjects of their own face on the screen is accompanied by an increase in activity of the cheek raiser and an increase in the proportion of time the gaze is fixed on the screen, while the perception of an empty screen – by a decrease in activity of the said muscle and a decrease in the proportion of time the gaze is fixed on the screen.

It can be noted that in the space of the first two latent structures, the perception of other face in all subjects is accompanied by an increase in the activity of the upper lip and chin raiser and the lip corner depressor.

According to Fig. 4, the third structure describes 10.1 % of the total variance and is determined by the activity of the nose wrinkler, chin raiser, gender and first trial features. Hence it follows that the perception of all the first samples is accompanied by an increase in activity of nose wrinkle and a decrease in activity of chin raiser in girls, and in boys – by a decrease in activity of nose wrinkle and an increase in activity of chin raiser.

The fourth structure describes 6.6 % of the total variance and is determined by the sign of the last trials, the activity of

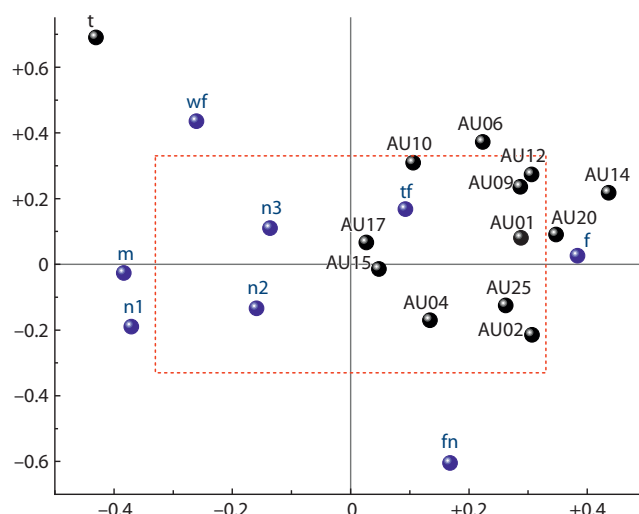


Fig. 3. Loads (correlation coefficients) of variables for latent structure No. 1 (X-axis; 53.8 % of total variance) and structure No. 2 (Y-axis; 14.9 % of total variance) in 2B-PLS model.

Here and in Fig. 4: black color – instrumental variables, blue color – feature variables (see the Table); inside the rectangle (red dotted line) the significance of the values of the correlation coefficients $p > 0.05$; markings of individual specificity are omitted to improve the readability of the graph.

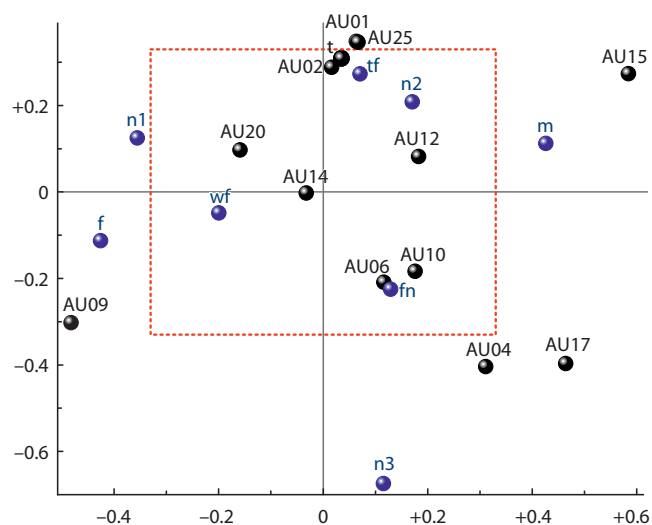


Fig. 4. Load (correlation coefficients) of variables for latent structure No. 3 (X-axis; 10.1 % of total variance) and structure No. 4 (Y-axis; 6.6 % of total variance) in 2B-PLS model.

the inner brow raiser, the depressor glabellae muscles, the chin raiser, and parted lips (see Fig. 4). What can be inferred from this is that the reaction to all third trials in all subjects is accompanied by an increase in the activity of the chin raiser and the depressor glabellae muscles, a decrease in activity of the inner eyebrow raiser and the degree of relaxation of the chin muscle and the circular muscle of the mouth, and parted lips.

It can be noted that in the space of latent structures No. 3 and 4, the perception of other face in all subjects is accompa-

nied by an increase in the proportion of the time of fixing the gaze on the screen and an increase in the activity of the inner and outer brow raiser, relaxation of the chin muscle and the circular muscle of the mouth, and parted lips.

Thus, in an EEG/ECG experiment, it is recommended for joint processing to include (apart from the influence of individual differences in implicit cognitive processes) the following covariate variables: gender, order of trials, presence of one's own face on the screen/blank screen.

Conclusions

Electroencephalogram is one of the most common methods for non-invasive study of the functional state of the human brain in healthy and clinical conditions. When analyzing the relationship between the EEG parameters and the behavioral activity of the subject, the motor (much less often verbal) responses of the subjects are usually chosen as behavioral metrics. This choice is primarily due to the fact that such responses are easy to mark in EEG recordings. We hypothesized that changes in the state of the facial muscles could serve as a behavioral phenotypic feature associated simultaneously with the personality characteristics of the survey participant, including their predisposition to mental disorders, and with endophenotypic parameters of brain rhythms.

In the present article, we propose a methodological idea for recording and processing facial video together with EEG recording. A pilot study was conducted aiming to find statistically significant covariates for facial expression to take into account in the analysis of EEG in the resting-state paradigm of functional rest and also when the subjects are being demonstrated a video recording of their own or someone else's face. This approach is based on the face muscles dynamics analysis of the subject on video, which is recorded simultaneously with the registration of EEG and ECG.

It was shown that the dynamics of facial muscle activity reflect controlled conditions that are not usually used in the analysis of EEG correlates of cognitive processes, but which, as follows from the results, may accompany certain implicit cognitive processes. Taking into account such covariates as the subject's gender, screen status (blank, own/other face) and sample number will increase the reliability of the assessment of the cognitive state of the subjects and provide additional information for interpreting the EEG/ECG results. The clustering of subjects by the factors of individual specificity of implicit cognitive processes will form a basis for effective profiling.

In the present study, we did not analyze EEG/ECG and psychometric data, as this is a pilot study with limited objectives. In the future, it is planned to increase the size of the experimental sample and conduct a more detailed comparison of the results of the analysis of the activity of the facial muscles with the results of other neurocognitive methods. For these promising tasks, we have worked out data obtaining methodology for profiling subjects according to the latent structures described by the authors, which allows to use the results of the generated model as additional variables for second-level summary models (including EEG, ECG data, etc.).

References

- Biswal B.B. Resting state fMRI: a personal history. *Neuroimage*. 2012; 62(2):938-944. DOI 10.1016/j.neuroimage.2012.01.090.
- Bringas-Vega M.L., Michel C.M., Saxena S., White T., Valdes-Sosa P.A. Neuroimaging and global health. *Neuroimage*. 2022;260:119458. DOI 10.1016/j.neuroimage.2022.119458.
- Engemann D.A., Mellot A., Hochenberger R., Banville H., Sabbagh D., Gemein L., Ball T., Gramfort A. A reusable benchmark of brain-age prediction from M/EEG resting-state signals. *Neuroimage*. 2022; 262:119521. DOI 10.1016/j.neuroimage.2022.119521.
- Greicius M.D., Flores B.H., Menon V., Glover G.H., Solvason H.B., Kenna H., Reiss A.L., Schlaggar A.F. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry*. 2007;62(5):429-437. DOI 10.1016/j.biopsych.2006.09.020.
- Kabbara A., Paban V., Weill A., Modolo Ju., Hassan M. Brain network dynamics correlate with personality traits. *Brain Connect*. 2020; 10(3):108-120. DOI 10.1089/brain.2019.0723.
- Khanin Yu.L. Quick Guide to C.D. Spielberger's Scale of State and Trait Anxiety. Leningrad, 1976. (in Russian)
- Knyazev G.G., Savostyanov A.N., Volf N.V., Liou M., Bocharov A.V. EEG correlates of spontaneous self-referential thoughts: a cross-cultural study. *Int. J. Psychophysiol*. 2012;86(2):173-181. DOI 10.1016/j.ijpsycho.2012.09.002.
- Kovaleva V.Yu., Pozdnyakov A.A., Litvinov Yu.N., Efimov V.M. Estimation of the congruence between morphogenetic and molecular-genetic modules of gray voles *Microtus s.l.* variability along a climatic gradient. *Ecol. Genet*. 2019;17(2):21-34. DOI 10.17816/ecogen17221-34.
- Li M., Wang Y., Lopez-Naranjo C., Hu S., Reyes R.C.G., Paz-Linares D., Areces-Gonzalez A., Hamid A.I.A., Evans A.C., Savostyanov A.N., Calzada-Reyes A., Villringer A., Tobon-Quintero C.A., Garcia-Agustin D., Yao D., Dong L., Aubert-Vazquez E., Reza F., Razzaq F.A., Omar H., Abdullah J.M., Galler J.R., Ochoa-Gomez J.F., Prichep L.S., Galan-Garcia L., Morales-Chacon L., Valdes-Sosa M.J., Tröndle M., Zulkifly M.F.M., Rahman M.R.B.A., Milakhina N.S., Langer N., Rudych P., Koenig T., Virues-Alba T.A., Lei X., Bringas-Vega M.L., Bosch-Bayard J.F., Valdes-Sosa P.A. Harmonized-Multinational qEEG norms (HarMNqEEG). *Neuroimage*. 2022;256:119190. DOI 10.1016/j.neuroimage.2022.119190.
- Milakhina N.S., Tamozhnikov S.S., Proshina E.A., Karpova A.G., Savostyanov A.N., Afonasiyeva E.B. Delta and gamma activity of resting-state EEG as one of the markers of risk of depressive disorders in migrants of subpolar and polar regions of Siberia. In: 2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB). Novosibirsk, 2020;90-92. DOI 10.1109/CSGB51356.2020.9214596.
- Nikolaeva E.I., Vergunov E.G. Evaluation of the relationship of facial expression asymmetry with inhibitory control and lateral preferences in physically active men. *Asimmetriya = Asymmetry*. 2021; 15(4):38-53. DOI 10.25692/ASY.2021.15.4.004. (in Russian)
- Polunin D., Shtager I., Efimov V. JACOBI4 software for multivariate analysis of biological data. *bioRxiv*. 2019;803684. DOI 10.1101/803684.
- Privodnova E.Yu., Slobodskaya H.R., Bocharov A.V., Saprygin A.E., Knyazev G.G. Default mode network connections supporting intra-individual variability in typically developing primary school children: An EEG study. *Neuropsychology*. 2020;34(7):811-823. DOI 10.1037/neu0000699.
- Proshina E.A., Savostyanov A.N., Bocharov A.V., Knyazev G.G. Effect of 5-HTTLPR on current source density, connectivity, and topological properties of resting state EEG networks. *Brain Res*. 2018; 1697:67-75. DOI 10.1016/j.brainres.2018.06.018.
- Rännér S., Lindgren F., Geladi P., Wold S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and

- algorithm. *J. Chemometrics*. 1994;8(2):111-125. DOI 10.1002/cem.1180080204.
- Rohlf F.J., Corti M. Use of two-block partial least-squares to study covariation in shape. *Syst. Biol.* 2000;49(4):740-753. DOI 10.1080/106351500750049806.
- Saprygin A., Lebedkin D., Savostyanov A., Vergounov E. Behavioral and neurophysiological study of subject's personality traits under recognition of sentences about self and others. In: *Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022)*. Abstracts the Thirteenth International Multiconference, Novosibirsk, 04–08 July 2022. Novosibirsk, 2022;950. DOI 10.18699/SBB-2022-556.
- Snyder A.Z., Raichle M.E. A brief history of the resting state: the Washington University perspective. *Neuroimage*. 2012;62(2):902-910. DOI 10.1016/j.neuroimage.2012.01.044.
- Volf N.V., Belousova L.V., Knyazev G.G., Kulikov A.V. Gender differences in association between serotonin transporter gene polymorphism and resting-state EEG activity. *Neuroscience*. 2015;284:513-521. DOI 10.1016/j.neuroscience.2014.10.030.

ORCID ID

A.N. Savostyanov orcid.org/0000-0002-3514-2901
E.G. Vergunov orcid.org/0000-0002-8352-5368
A.E. Saprygin orcid.org/0000-0001-6789-2953
D.A. Lebedkin orcid.org/0000-0002-4356-9067

Acknowledgements. Data collection and processing were supported by the Russian Science Foundation, grant No. 22-15-00142 "fMRI and EEG correlates of focus on oneself as a predisposition factor to affective disorders". The work of A.N. Savostyanov and A.E. Saprygin on data preprocessing was financed from the funds of the budget project of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences No. FWNR-2022-0020 "System biology and bioinformatics: reconstruction, analysis and modeling of the structural and functional organization and evolution of human, animal, plant and microorganism gene networks".

The authors thank V.E. Kalikin for the software implementation of the tool for analyzing AU by face image based on the OpenFace framework.

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

Received September 13, 2022. Revised November 17, 2022. Accepted November 17, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Development of a neural network for diagnosing the risk of depression according to the experimental data of the stop signal paradigm


M.O. Zelenskiy¹, A.E. Saprygin^{2, 3}, S.S. Tamozhnikov³, P.D. Rudych^{1, 3, 4}, D.A. Lebedkin^{1, 4}, A.N. Savostyanov^{1, 2, 3, 4} 

¹ Novosibirsk State University, Novosibirsk, Russia

² Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia

⁴ Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, Russia

 a-sav@mail.ru

Abstract. These days, the ability to predict the result of the development of the system is the guarantee of the successful functioning of the system. Improving the quality and volume of information, complicating its presentation, the need to detect hidden connections makes it ineffective, and most often impossible, to use classical statistical forecasting methods. Among the various forecasting methods, methods based on the use of artificial neural networks occupy a special place. The main objective of our work is to create a neural network that predicts the risk of depression in a person using data obtained using a motor control performance testing system. The stop-signal paradigm (SSP) is an experimental technique to assess a person's ability to activate deliberate movements or inhibit movements that have become inadequate to external conditions. In modern medicine, the SSP is most commonly used to diagnose movement disorders such as Parkinson's disease or the effects of stroke. We hypothesized that SSP could serve as a basis for detecting the risk of affective diseases, including depression. The neural network we are developing is supposed to combine such behavioral indicators as: the amount of missed responses, amount of correct responses, average time, the amount of correct inhibition of movements after stop-signal onset. Such a combination of indicators will provide increased accuracy in predicting the presence of depression in a person. The artificial neural network implemented in the work allows diagnosing the risk of depression on the basis of the data obtained in the stop-signal task. An architecture was developed and a system was implemented for testing motor control indicators in humans, then it was tested in real experiments. A comparison of neural network technologies and methods of mathematical statistics was carried out. A neural network was implemented to diagnose the risk of depression using stop-signal paradigm data. The efficiency of the neural network (in terms of accuracy) was demonstrated on data with an expert assessment for the presence of depression and data from the motor control testing system.

Key words: stop signal paradigm; artificial neural network; system for depression risk assessment; machine learning.

For citation: Zelenskiy 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 Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):773-779. DOI 10.18699/VJGB-22-93

Разработка нейронной сети для диагностики риска возникновения депрессии по экспериментальным данным стоп-сигнал парадигмы


М.О. Зеленских¹, А.Е. Сапрыгин^{2, 3}, С.С. Таможников³, П.Д. Рудыч^{1, 3, 4}, Д.А. Лебедин^{1, 4}, А.Н. Савостьянов^{1, 2, 3, 4} 

¹ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия

⁴ Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия

 a-sav@mail.ru

Аннотация. В настоящее время возможность спрогнозировать результат развития системы – залог успешного функционирования системы. Повышение качества и объема информации, усложнение ее представления, необходимость обнаруживать скрытые связи делают неэффективным, а чаще всего невозможным, применение классических статистических методов прогнозирования. Среди разнообразных методов прогнозирования особое место занимают методы, основанные на использовании искусственных нейронных сетей. Задачей нашей работы является создание нейронной сети, прогнозирующей риск возникновения депрессии у человека, с помощью данных, полученных при использовании системы тестирования показателей моторного контроля. Стоп-сигнал парадигма (ССП) – это экспериментальный метод, позволяющий оценить способность человека активировать целенаправленные движения или

подавлять движения, ставшие неадекватными внешним условиям. В современной медицине ССП чаще всего применяется для диагностики двигательных нарушений, таких как болезнь Паркинсона или последствия инсульта. Мы предполагаем, что ССП может служить основой для выявления риска развития аффективных заболеваний, включая депрессию. В разрабатываемой нами нейронной сети предполагается комбинирование таких поведенческих показателей, как количество пропущенных ответов, количество правильных ответов, среднее время, количество верных торможений после появления стоп-сигнала. Такой набор показателей обеспечит повышенную точность прогнозирования наличия депрессии у человека. Реализованная в работе искусственная нейронная сеть позволяет по данным, полученным с помощью фиксации реакции на стимулы со стоп-сигналом, диагностировать риск возникновения депрессии. Разработана архитектура и реализована система тестирования показателей моторного контроля у человека, затем протестирована в реальных экспериментах. Проведено сравнение нейросетевых технологий и методов математической статистики. Реализована нейронная сеть для диагностирования риска возникновения депрессии по данным ССП. На примере данных с экспертной оценкой на наличие депрессии и результатов, полученных при использовании системы тестирования показателей моторного контроля, продемонстрирована эффективность нейронной сети (с точки зрения точности).

Ключевые слова: стоп-сигнал парадигма; искусственная нейронная сеть; система тестирования; риск возникновения депрессии; машинное обучение.

Introduction

The ability to predict the result of the development of the system is the key to the successful functioning of the system. Improving the quality and volume of information, complicating its presentation, and the need to detect hidden connections makes it ineffective, and most often impossible, to use classical statistical forecasting methods. Among the various forecasting methods, methods based on the use of artificial neural networks occupy a special place.

The main objective of our work is to create a neural network that predicts the risk of depression in a person using data obtained using the motor control indicators testing system (Haykin, 2006). All data are taken from the open database of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences (ICBrainDB dataset <https://icbraindb.cytogen.ru/api-v2>).

A group of patients with depression was examined at the clinic of the Scientific Research Institute of Neurosciences and Medicine. The presence of major depressive disorder was diagnosed by a psychiatrist during a closed interview based on The International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) criteria. As a control group of healthy people, participants who had never been treated in psychiatric clinics and had not turned to psychiatrists for medical help were invited. All participants in the control group denied having any neurological or psychiatric diseases at the time of the examination or for five years before the examination. In addition, all the survey participants, both patients and control participants, denied the presence of alcohol or drug addiction and the usage of other psychoactive substances.

The main differences between artificial neural networks and methods of mathematical statistics are parallel processing of information and the ability to learn without a teacher, in other words, to self-study (<https://wiki.loginom.ru/articles/normalization.html>). Below, in the form of a table (Table 1), the results of comparing neural networks and methods of mathematical statistics according to the selected criteria are presented.

Resistance to noise is an important indicator when working with a large number of parameters and at the absence of explicit dependencies that we get from the data of the stop signal paradigm. Self-study makes it possible to perform tasks

without outside interference, which contributes to the search for patterns between parameters.

The use of mathematical statistics methods in the search for dependencies between the stop signal paradigm and the risks of depression cannot fully detect their presence due to the sensitivity of the methods to superfluous data, and even more so they cannot further predict the risk of depression in a person. Noise resistance and self-learning make usage of neural networks not simply preferable, compared to mathematical statistics, but necessary.

The neural network should accept a dataset consisting of data obtained using the stop signal paradigm as input and output the diagnostic result for the risk of depression.

The stop signal paradigm (SSP) is an experimental method that allows us to evaluate a person's ability to activate deliberate movements or suppress movements that have become inadequate to external conditions. In modern medicine, SSP is most often used to diagnose motor disorders, such as Parkinson's disease or the consequences of a stroke. We suggested that SSP can serve as a basis for identifying the risk of developing affective diseases, including depression. The neural network we are developing assumes a combination of behavioral indicators such as: the number of missed answers, the number of correct answers, the average time, the number of correct stops. Such a set of indicators will provide increased accuracy in predicting the presence of depression in a person.

The purpose of this work is to develop a neural network for predicting the risk of depression according to the stop signal paradigm. The artificial neural network implemented in the work makes it possible to predict the risk of depression based on the data obtained by registering the reaction to stimuli with a stop signal.

Materials and methods

Implementation of a neural network. The following table shows the technologies used for implementation along with a rationale (Table 2).

The architecture of the model. To work with the model and layers, the Sequential and Dense classes of the TensorFlow were used.

The Sequential class is a sequential neural network architecture, which is equivalent to sequential layer invocation (https://keras.io/api/layers/core_layers/dense/).

Table 1. Comparison of neural networks and mathematical statistics

Criteria	Neural networks	Methods of mathematical statistics
Saturation level	High saturation level	Low saturation level
Computing power	Require a lot of computing power	Require less computing power than artificial neural networks
Progression of algorithms	Continuous development of algorithms for building artificial neural networks	Development is slow
The absence of an unreasonable result	Presence of unreasonable results	Absence of unreasonable results
Time spent on development	A lot of development time	Less time and development costs
The amount of data to get the result	Requires a large amount of data for training	Needs less data than artificial neural networks
Resistance to noise	Resistant to noise	Not resistant to noise
An opportunity for self-learning	Availability of self-learning opportunities	Lack of self-learning opportunities

Table 2. Technology stack used

Technology	Rationale
Programming language: Python	At the moment, it allows easier and faster work with neural networks than other programming languages (e.g.: Java). Supports a wide range of libraries
Data Processing library: Pandas is an open-source library that provides tools for working with various data structures for the Python programming language (Vinogradova, 2012). The library was used for parsing experimental results and for further work with the dataset	Allows to process the data formats (comma- and tab-separated values)
Plotting library: matplotlib is a library for creating visualizations such as: histograms, bar charts, error bands, coherence graphs and much more (Ivanov et al., 2022). The library was used to plot the loss during training and validation of the neural network, the accuracy of training and validation	Selected for its capacity in constructing histograms
A library for interacting with artificial neural networks: Keras – a Python API for the TensorFlow (https://keras.io/about/)	One of the most popular neural network APIs
Version control system – Github	One of the most popular and easy to use version control systems

The Dense class implements the operation:

$$\text{output} = \text{activation}(\text{dot}(\text{input}, \text{kernel}) + \text{bias}), \quad (1)$$

where activation is the element-by-element activation function passed as an argument, kernel – is the matrix of all weights created by the layer, bias – is the displacement vector created by the layer (<https://keras.io/api/layers/activations/>).

Two layers were highlighted:

- layer x, that is, a layer for working with objects based on input data with the exception of the category of the test;
- layer y, that is, a layer for working with answers based on the category of the test subject.

To work with layer x, the relu activation function was used. The relu function returns a number if it accepts a positive argument, in other cases it returns 0 (<https://matplotlib.org/stable/index.html>). To work with layer y, the sigmoid activation function was used, which is necessary for probabilistic forecasting. Sigmoid activation function:

$$\text{sigmoid}(x) = \frac{1}{(1 + \exp(-x))}. \quad (2)$$

For small values, the function returns a value close to 0, and for large values, it returns close to 1, and the sigmoid always returns from 0 to 1 (https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php).

Data collection for training. In preparation for the development of the neural network, a balanced dataset was created based on data obtained during the examination of healthy people and patients with diagnosed depression.

The following inputs were highlighted:

- Missed – the number of missed responses from the test subject;
- Right – the total number of correct answers from the test subject;
- Av_time – average reaction time for the test subject during the experiment;
- Stop – the number of correct ignores on the stop signal of the test subject;
- Practice – the number of correct answers in the block “Practice” at the test;

Table 3. Selection of parameters on a balanced dataset

Trials	Validation	Training accuracy	Validation accuracy	Training losses	Validation losses	Conclusion
500	0.2	0.1–0.9	0.9–0	0.3–0	0.2–1	Does not satisfy
1000	0.2	1	0	0.2–0	0.3–1	Does not satisfy
200	0.2	0.1–0.9	0.9–0	0.3–0	0.2–0.8	Does not satisfy
100	0.2	1	0	0.13–0.3	0.45–0.8	Does not satisfy
1000	0.1	0.2–0.82	1–0	0.31–0	0.2–0.67	Does not satisfy
500	0.1	0.18–0.81	1–0	0.3–0	0.2–0.7	Does not satisfy
200	0.1	0.82	0	0.1–0	0.45–0.7	Does not satisfy
100	0.1	0.19–0.81	1–0	0.35–0	0.1–0.7	Does not satisfy
100	0.05	0.2–0.79	1–0	0.27–0.15	0.2–0.45	Does not satisfy
100	0.02	0.21–0.78	1–0	0.3–0.2	0.15–0.4	Does not satisfy

- Right_stop – the number of correct answers without taking into account the stop signal;
- Incor_stop – the number of incorrect reactions to the stop signal;
- Survive – the category of the test subject (healthy or diagnosed with depression).

Data preparation and normalization. Data normalization is a procedure for preprocessing input data, in which the values of the features forming the input vector are reduced to a specified range. Normalization is necessary because the initial values can vary over a large range and the operation of a neural network with such data can lead to an incorrect result (<https://keras.io/api/models/>). Normalization of data to the range [0...1] is important for setting a single privilege of features, in other words, for setting the same significance for each feature, which will allow them to be compared with each other in equal conditions.

All dataset inputs were selected for normalization, with the exception of Survive, since this parameter is an estimate and takes only two values: 0 or 1.

Network topology selection. Choosing the topology of an artificial neural network is one of the most important stages when using neural network technologies to solve practical problems. The adequacy of neural network model training directly depends on this stage (https://keras.io/api/models/model_training_apis/). Since we are faced with the task of classification and it is important to find any hidden connections, we need each artificial neuron to be connected to other neurons.

Based on the concepts of neural network types, a fully connected type was chosen, since, as mentioned earlier, each artificial neuron transmits its output to the rest of the neurons.

Experimental selection of training parameters. During this stage of neural network development, it is necessary to select optimal training parameters that will demonstrate the best accuracy and loss indicators. Selection is carried out by launching a neural network with possible parameters and a test dataset.

The following table (Table 3) shows the results of the experimental selection of training parameters, that is, the selection of the number of passes of the dataset from beginning to end (epochs) and the amount of data for validation (validation_split) on a balanced dataset (50 % healthy, 50 % with diagnosed depression, total 205).

Figure 1 demonstrates the accuracy of training and validation when training on a balanced dataset with a choice of epochs = 500 and validation_split = 0.2.

Thus, due to the lack of suitable parameters for further work, it was decided to use an unbalanced dataset (65 % of healthy, 35 % with diagnosed depression, only 500).

The following table shows the results of experimental selection of training parameters on an unbalanced dataset (Table 4).

Figure 2 demonstrates the accuracy of training and validation when training on an unbalanced dataset with epochs = 5000 and validation_split = 0.2.

Based on the results obtained, the number of passes from the beginning of the dataset to the end (epoch) = 4000 was selected, the amount of data for validation (validation_split) = 0.2.

Neural network training. To ensure the correctness of the artificial neural network, the sample was divided into two parts: training data for training, verification data for checking the operation of the neural network.

The compile and fit methods were used for training. The arguments of the compile method are: optimizer, loss function, metrics, loss weights, list of metrics. In the fit method, the arguments are: input data, target data, number of samples, number of epochs, list of callbacks, amount of data for validation (<https://pandas.pydata.org/pandas-docs/stable/>).

Arguments used in the compile method:

- loss = “mse” – root-mean-square error:

$$E[(X - \hat{X})^2] = E[(X - g(Y))^2], \quad (3)$$

let $\hat{X} = g(Y)$ be an estimate of a random variable, given the observation of a random variable Y (<https://www.journaldev.com/45330/relu-function-in-python/>);

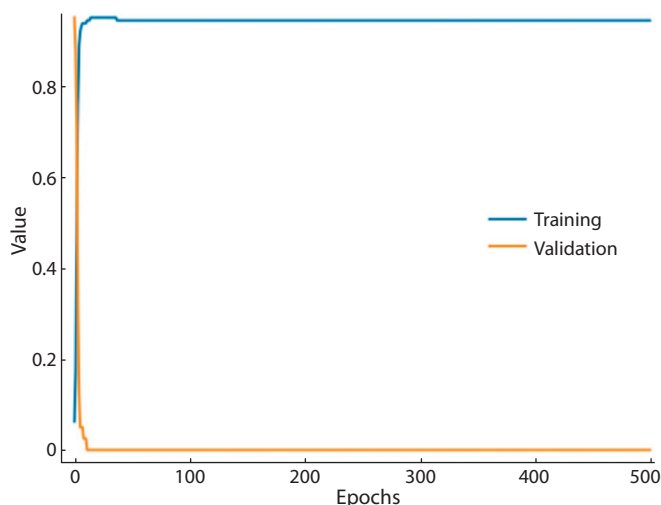


Fig. 1. An example of a graph of training accuracy and validation accuracy when training on a balanced dataset.

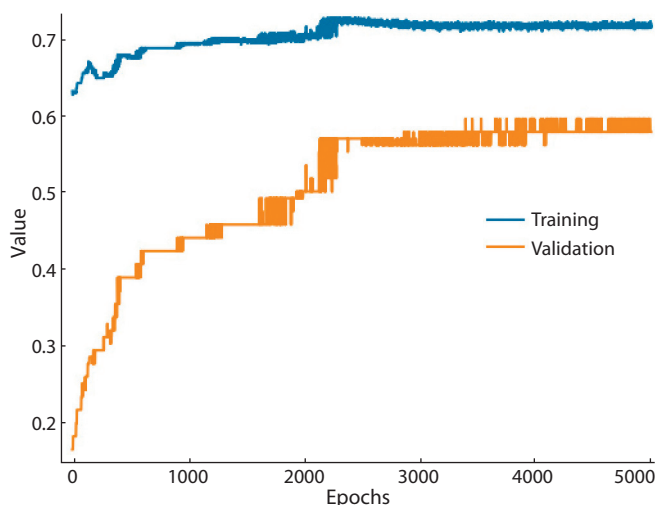


Fig. 2. An example of a graph of training accuracy and validation accuracy when training on an unbalanced dataset.

Table 4. Selection of parameters on an unbalanced dataset

Trials	Validation	Training accuracy	Validation accuracy	Training losses	Validation losses	Conclusion
1000	0.1	0.6–0.68	0.9–0.85	0.22–0.2	0.16–0.15	Does not satisfy
1000	0.05	0.9	0	0	0.9	Does not satisfy
1000	0.2	0.13–0.72	0–0.82	0.37–0.2	0.45–0.2	Satisfies
3000	0.2	0.13–0.7	0–0.8	0.37–0.19	0.45–0.2	Satisfies
4000	0.2	0.12–0.73	0–0.84	0.37–0.18	0.45–0.2	Satisfies

Table 5. Checking the adequacy of training

No.	1	2	3	4	5	6	7	8
Category	1	1	0	1	0	0	0	0
Result	0.767	0.824	0.24	0.927	0.316	0.293	0.276	0.367

Note. 0 – with diagnosed depression, 1 – without depression.

- optimizer = “sgd” – gradient descent optimizer taking into account momentum (<https://keras.io/api/optimizers/sgd/>);
 - metrics = [“accuracy”].
- Arguments used in the fit method:
- x – input data;
 - y – the target data, that is, the estimate;
 - epochs = “4000” – the number of epochs;
 - validation_split = “0.2” – the amount of validation data used in the training sample.

Checking the adequacy of training. Testing of the adequacy of training is carried out on data that were not in the training samples, in other words, new data for the neural network are used.

The following table (Table 5) shows an example of a sequence of values (PSurvived) obtained from the neural network, taking into account the category of data.

Results

Technical tests. For technical tests of the neural network, data from experiments on our system for testing human motor control indicators (without expert assessment for depression, that is, without clinical confirmation) were used, as well as previously unused data that did not participate in the training sample (with expert assessment).

The purpose of the technical tests is to study how the developed neural network will cope with the classification for the presence of risks of depression according to the stop signal paradigm.

Input data. The following input data were selected for the technical tests of the neural network:

- Unbalanced dataset (0.37 – with diagnosed depression, 0.63 – without depression);
- The maximum number of missed responses is 85;

Table 6. The results of the neural network

No.	Category of the subject	The result of the neural network	Evaluation based on the neural network result
1	Without an expert assessment for the presence of depression	0.8637	Healthy
2	»	0.5195	Healthy
3	»	0.6937	Healthy
4	»	0.7821	Healthy
5	»	0.7885	Healthy
6	»	0.4915	Presumed risk of depression
7	»	0.8123	Healthy
8	»	0.2868	Presumed risk of depression
9	Without an expert assessment for the presence of depression	0.7568	Healthy
10	Diagnosed with depression	0.1478	The risk of depression – corresponds to the category
11	Healthy	0.9487	Healthy – corresponds to the category
12	Diagnosed with depression	0.3227	The risk of depression – corresponds to the category
13	»	0.3114	The risk of depression – corresponds to the category
14	»	0.2721	The risk of depression – corresponds to the category
15	»	0.2993	The risk of depression – corresponds to the category

- The maximum total number of correct answers for the test – 92;
- The maximum average time per experiment for a test subject is 750.0;
- The maximum number of correct ignores for a stop signal from a test subject is 34;
- The maximum number of correct answers in the “Practice” block in the test – 31;
- The maximum number of correct answers without taking into account the stop signal is 65;
- The maximum number of incorrect reactions to the stop signal is 35;
- The amount of data for validation is 0.2, the number of epochs is 4000.

Test results. The following table describes the results of the neural network with an estimate of the values obtained (Table 6).

Thus, during the technical tests, the results of the neural network were obtained, which demonstrate which category (healthy/at risk of depression) the test subject belongs to. The obtained indicators fully correspond to the diagnoses.

Conclusion

Based on the experimental data obtained using the stop signal paradigm, a dataset was formed. The implementation of a neural network for diagnosing the risk of depression according to the stop signal paradigm has been developed and further tested. Using the example of data with an expert assessment for the presence of depression and data obtained using the motor control indicators testing system, the accuracy of the

neural network classification was shown. The test results in the form of performance indicators of the neural network are described below:

Indicator	Meaning
Training losses	0.1657
Training accuracy	0.7821
Validation losses	0.2415
Validation accuracy	0.6667

The stop signal paradigm is commonly used to diagnose motor disorders such as Parkinson’s disease, childhood hyperactivity or post-traumatic disorders. Previously, the stop signal paradigm was not used by anyone to diagnose depression. We applied this technique in combination with neural network methods and showed that the results of SSP make it possible to efficiently classify people into patients with depression and people without depression. It should also be noted that we did not compare patients with depression with patients with other non-depression-related neurological diseases. Therefore, at the moment, it is not yet clear whether our method allows us to divide patients with different disorders into different subclasses.

References

About Keras [Electronic resource]. URL: <https://keras.io/about/>.
Dense layer [Electronic resource]. URL: https://keras.io/api/layers/core_layers/dense/.
Haykin S. Neural Networks. A Comprehensive Foundation. Moscow: Williams Publ., 2006. (in Russian)
Ivanov R., Kazantsev F., Zavarzin E., Klimenko A., Milakhina N., Matushkin Yu., Savostyanov A., Lashin S. ICBrainDB: An integrated

- database for finding associations between genetic factors and EEG markers of depressive disorders. *J. Pers. Med.* 2022;12(1):53. DOI 10.3390/jpm12010053.
- Layer activation functions [Electronic resource]. URL: <https://keras.io/api/layers/activations/>.
- Matplotlib documentation – Matplotlib 3.5.1 documentation [Electronic resource]. URL: <https://matplotlib.org/stable/index.html>.
- Mean Squared Error (MSE) [Electronic resource]. URL: https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php.
- Model training APIs [Electronic resource]. URL: https://keras.io/api/models/model_training_apis/.
- Models API [Electronic resource]. URL: <https://keras.io/api/models/>.
- Normalization of input vectors (Normalization) – Loginom Wiki [Electronic resource]. URL: <https://wiki.loginom.ru/articles/normalization.html>.
- Pandas documentation – pandas 1.4.2 documentation [Electronic resource]. URL: <https://pandas.pydata.org/pandas-docs/stable/>.
- ReLU Function in Python – JournalDev [Electronic resource]. URL: <https://www.journaldev.com/45330/relu-function-in-python>.
- SGD [Electronic resource]. URL: <https://keras.io/api/optimizers/sgd/>.
- Vinogradova E.Yu. Principles of choosing the optimal topology of neural network to support managerial decision making. *Upravlenets = The Manager*. 2012;7-8:74-78. (in Russian)

ORCID ID

A.E. Saprygin orcid.org/0000-0001-6789-2953
P.D. Rudych orcid.org/0000-0003-3105-6931
D.A. Lebedkin orcid.org/0000-0002-4356-9067
A.N. Savostyanov orcid.org/0000-0002-3514-2901


Acknowledgements. Data processing using a neural network was carried out with the financial support of a grant from the Russian Science Foundation, No. 22-75-10105. The preparation of the experimental data base was carried out with the participation of A.E. Saprygin and A.N. Savostyanov within the framework of the budget project of Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences No. FWNr-2022-0020 "System biology and bioinformatics: reconstruction, analysis and modeling of the structural and functional organization and evolution of human, animal, plant and microbial gene networks".

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

Received September 19, 2022. Revised November 20, 2022. Accepted November 21, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

A software system for modeling evolution in a population of organisms with vision, interacting with each other in 3D simulator

A.P. Devyaterikov^{1, 2}, A.Yu. Palyanov¹ 

¹ A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia


 palyanov@iis.nsk.su

Abstract. Development of computer models imitating the work of the nervous systems of living organisms, taking into account their morphology and electrophysiology, is one of the important and promising branches of computational neurobiology. It is often sought to model not only the nervous system, but also the body, muscles, sensory systems, and a virtual three-dimensional physical environment in which the behavior of an organism can be observed and which provides its sensory systems with adequate data streams that change in response to the movement of the organism. For a system of hundreds or thousands of neurons, one can still hope to determine the necessary parameters and get the functioning of the nervous system more or less similar to that of a living organism – as, for example, in a recent work on the modeling of the *Xenopus* tadpole. However, of greatest interest, both practical and fundamental, are organisms that have vision, a more complex nervous system, and, accordingly, significantly more advanced cognitive abilities. Determining the structure and parameters of the nervous systems of such organisms is an extremely difficult task. Moreover, at the cellular level they change over time, these including changes under the influence of the streams of sensory signals they perceive and the life experience gained, including the consequences of their own actions under certain circumstances. Knowing the structure of the nervous system and the number of nerve cells forming it, at least approximately, one can try to optimize the initial parameters of the model through artificial evolution, during which virtual organisms will interact and survive, each under the control of its own version of the nervous system. In addition, in principle, the rules by which the brain changes during the life of the organism can also evolve. This work is devoted to the development of a neuroevolutionary simulator capable of performing simultaneous functioning of virtual organisms that have a visual system and are able to interact with each other. The amount of computational resources required for the operation of models of the physical body of an organism, the nervous system and the virtual environment was estimated, and the performance of the simulator on a modern desktop computing system was determined depending on the number of simultaneously simulated organisms.

Key words: nervous system; vision system; virtual organism; population; computational modeling; neuroevolution simulator.


For citation: Devyaterikov A.P., Palyanov A.Yu. A software system for modeling evolution in a population of organisms with vision, interacting with each other in 3D simulator. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):780-786. DOI 10.18699/VJGB-22-94

Программная система на основе 3D симулятора для моделирования эволюции в популяции организмов, обладающих зрительной системой

А.П. Девятериков^{1, 2}, А.Ю. Пальянов¹ 

¹ Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

 palyanov@iis.nsk.su

Аннотация. Создание компьютерных моделей, имитирующих работу нервных систем живых организмов с учетом их морфологии и электрофизиологии, – один из важных и перспективных разделов вычислительной нейробиологии. При наличии возможности стремятся моделировать не только нервную систему, но и тело, мышцы, сенсорные системы и виртуальную трехмерную физическую среду, в которой можно наблюдать поведение организма и которая обеспечивает его сенсорные системы адекватными потоками данных, изменяющимися в ответ на движение организма. Для системы из сотен или тысяч нейронов еще можно надеяться задать необходимые параметры и получить функционирование нервной системы, более-менее сходное с таковым для живого организма, как, например, в недавней работе по моделированию головастика *Xenopus*. Однако наибольший инте-

рес, как практический, так и фундаментальный, представляют организмы, обладающие зрением, более сложной нервной системой и, соответственно, значительно более развитыми когнитивными способностями. Определить структуру и параметры нервных систем таких организмов представляется исключительно сложной задачей. Более того, они изменяются с течением времени, в том числе под воздействием воспринимаемых ими потоков сенсорных сигналов и полученного жизненного опыта, включая последствия собственных действий при тех или иных обстоятельствах. Зная структуру нервной системы и число образующих ее нервных клеток хотя бы приблизительно, можно попытаться оптимизировать начальные параметры модели посредством искусственной эволюции, в процессе которой виртуальные организмы будут взаимодействовать и выживать – каждый под управлением собственной версии нервной системы. Помимо этого, эволюционировать могут и правила, по которым мозг изменяется на протяжении жизни организма. Данная работа посвящена созданию нейроэволюционного симулятора, способного осуществлять одновременное функционирование виртуальных организмов, обладающих зрительной системой, которые взаимодействуют между собой. Приведены расчеты, показывающие, сколько вычислительных ресурсов требуется для работы моделей физического тела организма, нервной системы и виртуальной среды обитания, а также определена производительность симулятора на современной настольной вычислительной системе в зависимости от числа одновременно моделируемых организмов.

Ключевые слова: нервная система; зрительная система; виртуальный организм; популяция; компьютерное моделирование; нейроэволюционный симулятор.

Introduction

Computational models imitating the functioning of living organisms' nervous systems, based on their electrophysiological and morphological data, are powerful tools in neuroscience. With their help it is possible, on the basis of knowledge and ideas about the functioning of individual nerve cells and the mechanisms of interaction between them, to calculate the dynamics of the activity of networks of nerve cells. The model of the nervous system functioning in combination with the model of the body of an organism equipped with muscular and sensory systems, placed in a virtual three-dimensional physical environment, provides the researcher with significant advantages. First, one can observe and register both the behavior of the body model of an organism and the activity of the nervous system, up to the activity of individual nerve cells, their processes and synapses. Secondly, the model of the nervous system receives a stream of signals from the virtual environment that change in response to the actions of the organism, driven by a muscular system controlled by its "brain", i. e. there is a constant feedback between actions and their consequences, just like in reality. One of the goals of such modeling is to check the adequacy of neural network models by comparing the activity of nervous systems of a real organism and its virtual 'twin', as well as their behavior.

Probably the most well-known creature in this context is one of the most simple multicellular organisms, invertebrate *Caenorhabditis elegans*, whose nervous system is composed of just 302 neurons (Sarma et al., 2018). Also, sufficiently convincing similarity between the real organism and the model was achieved for the *Xenopus* frog tadpole at the two-day stage of development, whose nervous system model was represented by a neural network composed of approximately 2300 neurons (Ferrario et al., 2021). However, neither *C. elegans*, nor the two days old *Xenopus* tadpole has a visual system.

Attempts to model much more complex organisms such as a mouse (~70 million neurons (Herculano-Houzel et al., 2006)) or a rat (~200 million neurons (Herculano-Houzel, Lent, 2005)), including their nervous systems, have also been made. However, to date, their virtual twins have not yet been created. The work aimed at reverse engineering and modeling the nervous system of the *Drosophila* fruit fly (~100 thousand

neurons (Scheffer et al., 2020)) is also in progress. Another extremely promising object of investigation and modeling is ants (~250 thousand neurons (Moffet et al., 2021)). These insects have immobile compound eyes, consisting of 100...3000 ommatidia – structural and functional units of such eyes (their number depends on the type of ant and its specialization), providing color vision with a rather modest resolution (from 10×10 to 55×55 "pixels"). Thus, for example, the eyes of *Myrmica ruginodis* usually have 109 to 169 ommatidia, and those of *Camponotus crassus* and *Pseudomyrmex adustus*, which are active during daylight hours – up to 700 and 930, correspondingly (Aksoy, Camlitepe, 2018), and the maximal known number of ant ommatidia per eye, near 3000, was registered in tropical species *Gigantiops destructor* (Macquart et al., 2006).

It is noteworthy that ants are the simplest organisms that successfully pass the mirror test, i. e. they are able to distinguish their own reflection in a mirror from another ant, which they can see through ordinary transparent, non-mirror glass of the same size (Cammaerts M.-C., Cammaerts R., 2015). The principle of conducting a mirror test is worth mentioning. In front of a mirror, ants clean themselves up or make unusual movements of their head and antennae, which is not observed when they see relatives behind the glass. If a small mark (e. g. blue) is applied on the front of an ant's head, then when it sees itself in the mirror, it will try to get rid of it, try to clean it off with the help of its legs. And if the mark is the same color as the body of the ant, or if it is applied to the back of the head, not visible in the mirror, then the ant will not show concern and attempts to clean it off. Thus, the ants notice the mark on themselves and behave as if they understand that it is on themselves, and not on another ant, relying solely on visual signals.

Computational modeling of both a single ant, with or without a mark, able to see itself in a mirror, as well as multiple ants that can see and interact with each other and with surrounding objects is of considerable scientific interest. Orientation on the terrain in ants is also carried out mainly through vision (Buehlmann et al., 2020).

What are the requirements for a software system and computing hardware capable of performing computer simulation

of a group of virtual organisms imitating ants (including body, muscle, sensory and nervous systems) and their habitat? It is assumed that organisms can interact with each other in the physical world and “see” each other, i.e. their nervous system receives a stream of video data corresponding to the first-person view as input. The problem of “digitizing” the structure of the nervous system, including 3D morphology of each neuron, its processes and synapses, is extremely labor- and time-consuming. However, this may not be essential, since the brain, even in ants, is quite plastic and undergoes noticeable structural changes during the life of the organism (Penick et al., 2021). At the same time, not much is known about the mechanisms underlying brain changes throughout life at the level of single neurons and connections between them. Therefore, it makes sense to pose the problem of modeling an organism that has the body and sensory systems of an ant (at least visual and mechanosensory, as well as the simplest olfactory and taste receptors) and a nervous system with a similar number of neurons and synapses, but without a fixed connectome. How fast can such modeling be carried out and can one expect that virtual evolution in such a system will help artificial neural networks to achieve cognitive capabilities that will allow virtual organisms to effectively survive, solving more or less complex tasks related to finding food, avoiding hazards and performing other activities?

Materials and methods

Software system. In accordance with the subject of the article, we are using computational modeling to deal with the problems to be solved – the research is carried out based on the software that we developed for conducting numerical experiments in the field of neuroevolutionary modeling. It is based on a modern 3D physics engine named Unigine (unigine.com), which is used for developing games, virtual reality systems, interactive visualization software, educational systems in various areas, etc., supporting Windows and Linux platforms.

The physics simulation module supports collision detection, rigid body physics, various types of joints (hinged, ball, prismatic, cylindrical, etc.), dynamic destruction of objects, cloth, floating objects, force fields, time reversal, etc. (<https://developer.unigine.com/ru/docs/latest/principles/physics/>). In Unigine it is possible to use mirrors, which may be useful in the future for conducting a “mirror test”. Also, it has built-in C++ programming language, which allows to develop and use one’s own program code – for example, to model networks of neurons that receive signals from virtual organisms sensory systems and control their movements.

An “ant” body model. The simple “ant” body model that we designed and used as a first prototype to evaluate the performance of the simulator is shown in Figure 1. In the future, it is planned to develop and use a much more detailed and realistic version.

In the simplest test scene, food particles (shown in green) and several dozen virtual organisms are randomly placed on the plane (Fig. 2).

Visual system. Figure 3 shows examples of images perceived by a “video camera” located on the body’s head, which is directed forward (at the moment only color mono-vision is implemented, although stereo is also planned for the future). The resolution of frames of ant’s video stream was chosen to

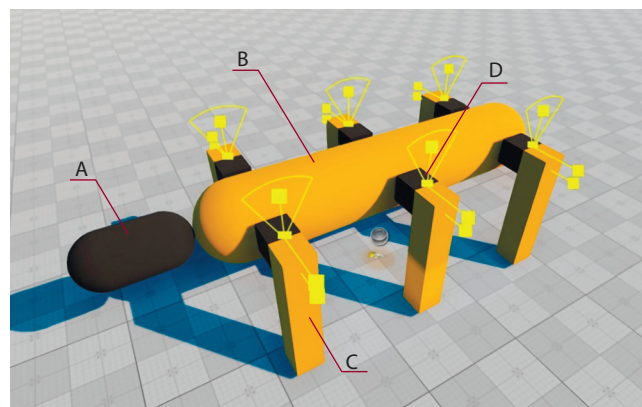


Fig. 1. Simple “ant” 3D body model, general view.

A – head, B – body, C – legs, D – a joint connecting body and legs. The head has a movable connection with the body.

be 30×30 , which approximately corresponds to the average spatial resolution of visual systems of real ants considered earlier. Since the images themselves are quite small, for the convenience of perception in the figure they are proportionally enlarged by 5 times (one color square of 5×5 pixels corresponds to one real “receptor” pixel).

An image can be represented as three matrices, each of which represents a separate color channel (red – R, green – G and blue – B). Each matrix has a size of 30×30 , forming an array of data, *Input*, consisting of 2700 elements, organized in the following way:

$$\begin{aligned} \text{Input}(r) &= R(i, j), \quad r = i \cdot 30 + j, \\ \text{Input}(g) &= G(i, j), \quad g = i \cdot 30 + j + 900, \\ \text{Input}(b) &= B(i, j), \quad b = i \cdot 30 + j + 1800, \\ &\text{where } 0 \leq i < 30, \quad 0 \leq j < 30. \end{aligned}$$

The simulation has a certain frame refresh rate, depending on the computational performance of the hardware, the complexity of the simulated scene and the number of “ants”. With a certain frequency, each individual forms such an array, the content of which enters the “nervous system” of the organism.

Nervous system. Visual signals enter “nervous systems” of virtual organisms, which at the very beginning of the simulation, for the first generation of “ants”, are randomly generated networks of artificial neurons, similar to those used in perceptrons (Rosenblatt, 1962) for recognition of letters, digits and geometrical figures. In our case, the number of neurons in each network was about 3000. Within the lifetime of one individual, networks have a static topology. Perceptron consists of S-elements (sensory), one or more layers of A-elements (associative) and R-elements (reacting). A-elements are defined by a set of weight matrices A_1, A_2, \dots, A_n and bias vectors b_1, b_2, \dots, b_n . The array *Input*, mentioned above, is processed in the following way:

$$\text{result}_i = A_i \cdot \text{result}_{i-1} + b_i,$$

where result_0 is a layer of sensory elements, containing an array of visual data perceived by an “ant”, and $i = 1, \dots, n$. And activation of R-elements as a result of visual data processing leads to the corresponding actions performed by the ant (change of speed, turn to the left or to the right).

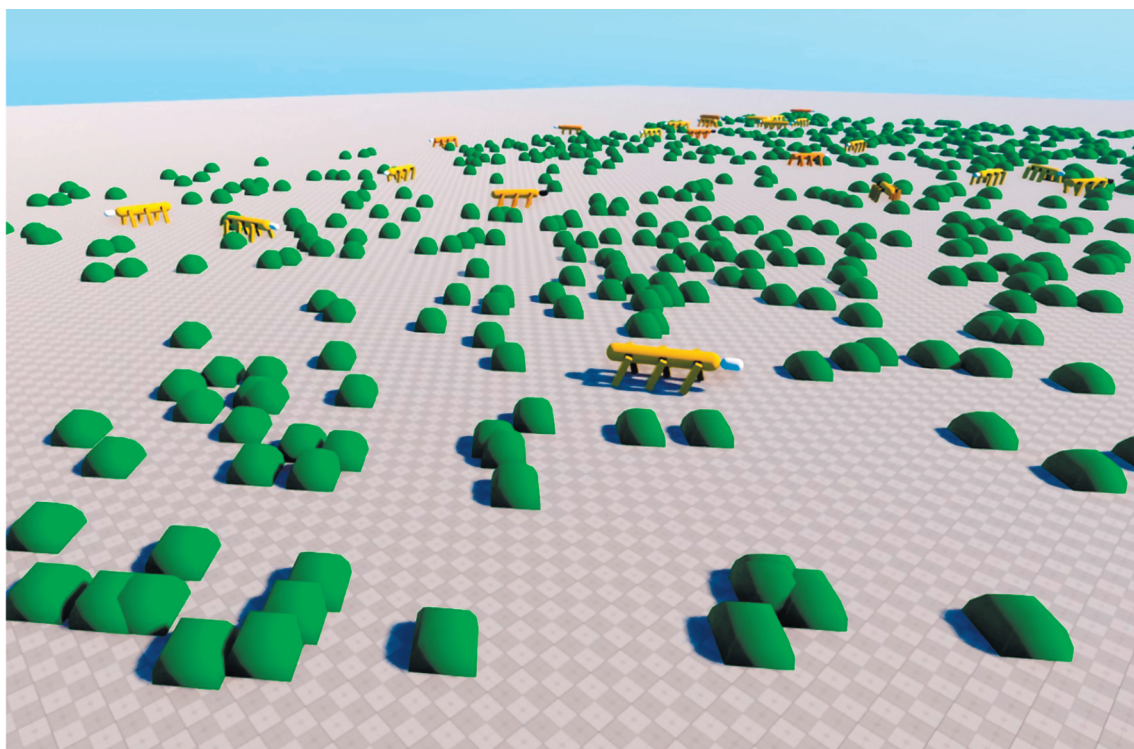


Fig. 2. General view of the simulation – test scene with a few dozens of virtual organisms.

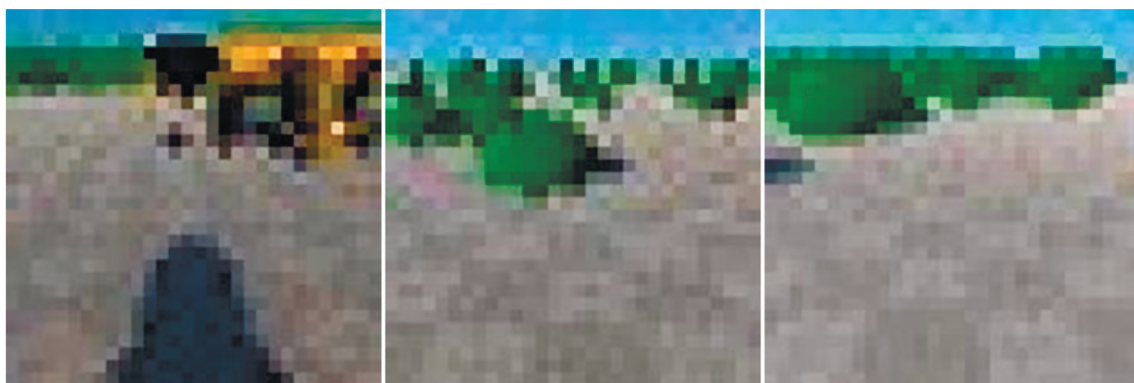


Fig. 3. A few examples of the “first-person view”.

In the first one (on the left), one can distinguish another individual (top, in brown tones) and the shadow of the virtual organism perceiving this image (dark gray).

Simulation of evolution. Some variants of weights matrices of perceptrons described above provide more efficient survival, i. e. the ability to perceive “first-person view” visual signals, analyze them and control the movement of the body in such a way that an organism regularly reaches food particles and maintains the necessary “energy level” in the body (satiated state). Organisms that remain hungry for too long die out and the “long-livers” have the opportunity to generate offsprings that inherit the structure of their neural networks. Currently, offspring is generated by only one parent (in nature such a reproduction mechanism, called parthenogenesis, also exists – in many types of arthropods, including 8 species of ants, as well as in about 70 species of vertebrates).

In the simulator, the current “energy level” of the organism is indicated as $Satiety(t)$, with which the following quantities are associated:

$MaxSatiety$ – maximum organism satiety (15 by default);

$BirthSatiety = MaxSatiety \cdot 0.7$ – the satiety of the organism, upon reaching which it gives birth to a descendant. When it happens, half of the available resources remains with the organism, and half passes to the descendant.

Each organism is initialized with $Satiety(0) = 8$. Each time after a certain period, it loses one satiety unit (because organism functioning “consumes energy”). At $Satiety(t) = 0$ the organism dies. When eating food, the organism gains a satiety point until $MaxSatiety$ is reached.

The child inherits the parent’s neural network with changes that are carried out according to the following rules:

- ε, δ – random values which are distributed uniformly;
- $\varepsilon \in [a, b]$ – probability of changes in neuron parameters (“mutation”), $0 \leq a \leq b \leq 1$;
- $\delta \in [c, d]$ – the amount of weight change in the matrix element as a result of mutation, $c \leq d$. Parameters a, b, c and d can be changed by user.

Every element of weight matrices and bias vectors, $A_k(i, j)$ and $b_k(l)$ ($k = 1, \dots, n$) changes by $+\delta$ or $-\delta$ with probability ε .

Results

At the current stage of the work, the main achieved result is the development of the simulator prototype (including a three-dimensional physical world, a model of the physical body of an ant, a model of the visual system and a model of the nervous system), as well as measurements of its performance on various computing systems, depending on their characteristics and on the number of neurons in the nervous system of virtual organisms. The source code of the simulator is available in the following repository (<https://github.com/NotNa19/AntPrototype>). Perspectives of further development of this project depend on the ability to perform neuroevolutionary modeling for at least one, but preferably for more virtual organisms, whose “nervous systems” are comparable to those of real ants in terms of the number of nerve cells.

Table 1 contains the characteristics of the computational hardware used in the testing and the maximum number of virtual organisms modelled simultaneously for which the simulation still remains stable. In this case, “stable work” means the correct functioning of organisms and their physical bodies. The fact is that in the current version of Unigine, at a low frame rate, delays between the movement of various components of the organism may occur, the processing of collisions between the objects, including “organisms” and “food”, may not always work correctly, and some other problems of this kind may happen as well. It is possible to fix these problems and it is planned for the future, but it requires a deeper knowledge about the mechanisms of the 3D engine. With a screen resolution of 1920×1080 pixels and its refresh rate (frames per second, FPS) of at least 30 per second, the simulator remains stable. However, the number of individuals simulated at the same time affects the performance. The following values were obtained on our computational hardware:

Table 1. The maximum size of the population of virtual organisms at which the simulator is stable, depending on the characteristics of the hardware used

Characteristics of the computing system	The maximum number of virtual organisms at which the simulator is stable
CPU Intel Core i5-7300HQ 2.50 GHz GPU GeForce GTX1050 Ti, 4 Gb	50
CPU AMD Ryzen 7 2700X 3.70 GHz GPU NVIDIA GeForce 1060, 6 Gb	80
CPU AMD Ryzen 5 5600X 3.7/4.6 GHz GPU MSI GeForce RTX 3060 Ti, 8 Gb	150

Detailing of the time spent on various stages of the simulation showed that with a small size of nervous systems (thousands to tens of thousands of neurons), the most significant factor limiting the speed of its operation is the process of obtaining “first-person view” video stream data for the ant population, even considering the fact that the multithreading of calculations is provided by the engine itself. Dependence of the maximum number of individuals in the simulation on the number of neurons in the “nervous system” of the virtual organism (all individuals have the same number) has also been investigated. The following values were obtained for GeForce RTX 3060 Ti + AMD Ryzen 5 5600X (Table 2).

Table 2. The maximum population size of virtual organisms at which the simulator is stable, depending on the number of neurons in their “nervous system”

The number of neurons	The maximum number of virtual organisms at which the simulator is stable
3000	150
10 000	50
100 000	10

The costs of 3D scene visualization for an external observer also have a noticeable impact on the performance of the system. Measurements performed at the computational system composed of AMD Ryzen 7 2700X 3.70 GHz CPU and NVIDIA GeForce 1060 6 Gb GPU revealed the following:

- When performing a simulation with an empty scene (with or without visualization for an external observer), stable 9000 clock cycles in 60 seconds (an average of 150 clock cycles/sec) are obtained.
- When performing a simulation with 80 organisms, with visualization for an external observer, we get 5400 cycles in 60 seconds (an average of 90 cycles/sec), and 7800 cycles in 60 seconds (an average of 130 cycles/sec) without visualization.
- With a higher load (100 individuals and more food), we obtained 1800 cycles in 60 seconds with visualization (on average 30 cycles/sec) and 4500 clock cycles in 60 seconds without visualization (an average of 75 clock cycles/sec).

Thus, visualization for an external observer (user) plays a fairly significant role in the overall performance of the system and thus it makes sense to turn it on only when it is really necessary – for example, in cases of debugging or recording demo videos illustrating the functioning of the simulator.

The work of the genetic algorithm can be illustrated by the dependence of the individuals’ lifetime, which increases as the number of generations grows. The curves shown in Figure 4 were obtained based on 10 runs of the simulator with the same parameters.

It can be seen that over time there are individuals appearing in the population whose lifetime is many times longer than the lifetime of individuals with randomly generated neural network parameters that have not yet passed natural selection. At the behavior level and with visual observation, it is expressed in the fact that the most adapted virtual organisms

purposefully move towards the particles of “food” and avoid moving away from the central area of space with the largest concentration of “food”, i. e. they are successfully adapted to their living conditions.

Discussion

The current neural network architecture is quite simple and at this stage was used mainly for testing the system as a whole and for evaluating its performance at an early stage of development. Currently, the following much more advanced and modern neural network architecture, which is a combination of a convolutional neural network (LeCun, Bengio, 1995) (for working with incoming video data) and the NEAT algorithm (NEuroevolution of Augmenting Topologies) (Stanley, Miikkulainen, 2002) is being implemented. NEAT can change not only the weight parameters, but also the structure of the neural network during the lifetime of the organism. The convolutional neural network will transform the details of the image to some abstractions, and the NEAT algorithm will handle the behavioral part of the virtual organism and work with the results of the functioning of this convolutional neural network.

In addition to this variant, self-organizing networks such as neocognitron (Kunihiko, 1980) are quite promising in terms of architecture as well. There are also neural networks that are much more realistic in terms of electrophysiology and neuromorphology. They are based on the Hodgkin–Huxley nervous cell model (Hodgkin, Huxley, 1952), in which it is represented in the form of compartments characterized by electrical capacitances and resistances, with calculations of membrane potentials and ion currents. The modern implementation of this model with support of parallel computing on GPUs has the following performance indicators. In the work (Stimberg et al., 2020), a neural network of 64 thousand neurons required about 0.6 sec of working time on a Tesla V100 GPU (with a performance of 14.1 TFLOPS in FP32 mode) to calculate 1 sec of simulation time (i. e. real time), and about 3 sec of calculations per 1 sec of simulation time – for neural networks of 256 thousand neurons. At the same time, numerical integration of the equations describing the system occurs with a time interval not exceeding 0.1 msec to ensure the accuracy of calculations and stability of the system, and each neuron on average has about 1000 connections (80 % of which are activating, and 20 % are inhibiting).

Recently, the research on new neural network architectures has been quite actively conducted, and many of obtained results have been successfully applied in practice. Particularly, in the field of neuroevolutionary methods, quite a wide range of promising variants has been considered, classified and compared in the dissertation (Khlopko, 2016, Ch. 1) and in the review article (Ma, Xie, 2022). In the future we plan to implement the most suitable and promising of them in the presented simulator and explore the limits of their “cognitive capabilities” while controlling the virtual “ants”.

Conclusion

Modern GPUs, such as, for example, NVidia 3080 Ti, with 10240 parallel CUDA computing cores, have a performance of 34.1 TFLOPS, and the upcoming 4080 Ti is expected to have 67.6 TFLOPS. Thus, the technological capability to simulate

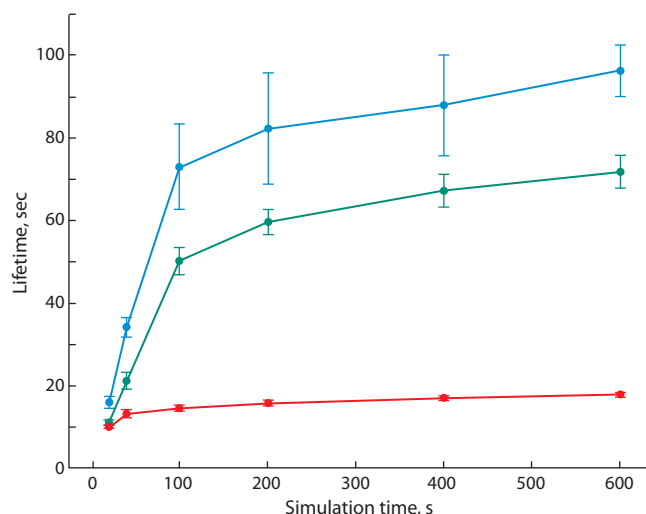


Fig. 4. The dependence of the maximum lifetime of an individual from the population for the entire period from the beginning to the present moment (blue curve), at the moment (green curve), and during the average lifetime of the population (red curve), indicating the root-mean-square deviation.

The data is obtained from 10 simulation runs.

a single virtual organism with a biologically realistic neural network of 256 thousand neurons and 256 million connections between them, with a numerical integration time step equal to 0.1 msec, on a single GPU, has already been achieved. It is comparable to the neural network of the real ant’s nervous system, which includes about 250 thousand neurons.

Our calculations for virtual organisms with neural networks of several thousand elements have shown that the computational costs of neural networks and the virtual physical environment are relatively small, and the main limiting factor for the system performance is video data streams in the “first person view” mode, carrying visual information. However, in the case of neural networks consisting of hundreds of thousands of neurons, the “nervous system” becomes the main consumer of computing resources. Thus, given the above, a modern desktop computing system with a powerful modern GPU has enough performance to provide a real time simulation of a virtual organism with a “nervous system” based on the Hodgkin–Huxley model, with a number of neurons composing its nervous system equivalent to that of a real ant. And if there are multiple GPUs in one workstation, the number of simultaneously simulated ants interacting with each other can be increased in proportion to the number of GPUs.

References

- Aksoy V., Camlitepe Y. Spectral sensitivities of ants – a review. *Anim. Biol.* 2018;68(1):55-73. DOI 10.1163/15707563-17000119.
- Buehlmann C., Wozniak B., Goulard R., Webb B., Graham P., Niven J.E. Mushroom bodies are required for learned visual navigation, but not for innate visual behavior, in ants. *Curr. Biol.* 2020; 30(17):3438-3443.e2. DOI 10.1016/j.cub.2020.07.013.
- Cammaerts M.-C., Cammaerts R. Are ants (Hymenoptera, Formicidae) capable of self recognition? *J. Sci.* 2015;5(7):521-532.
- Ferrario A., Palyanov A., Koutsikou S., Li W., Soffe S., Roberts A., Borisyuk R. From decision to action: detailed modelling of frog tadpoles reveals neuronal mechanisms of decision-making and repro-

- duces unpredictable swimming movements in response to sensory signals. *PLoS Comput. Biol.* 2021;17(12):e1009654. DOI 10.1371/journal.pcbi.1009654.
- Herculano-Houzel S., Lent R. Isotropic fractionator: a simple, rapid method for the quantification of total cell and neuron numbers in the brain. *J. Neurosci.* 2005;25(10):2518-2521. DOI 10.1523/jneurosci.4526-04.2005.
- Herculano-Houzel S., Mota B., Lent R. Cellular scaling rules for rodent brains. *Proc. Natl. Acad. Sci. USA.* 2006;103(32):12138-12143. DOI 10.1073/pnas.0604911103.
- Hodgkin A.L., Huxley A.F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 1952;117(4):500-544. DOI 10.1113/jphysiol.1952.sp004764.
- Khlopikova O.A. Methods and algorithms for the intellectualization of decision making under conditions of indeterminacy based on neural networks and evolutionary modeling. Ph.D. Thesis. Moscow, 2016. (in Russian)
- Kuniyoshi F. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernetics.* 1980;36(4):193-202. DOI 10.1007/BF00344251.
- LeCun Y., Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib M.A. (Ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge; London: Bradford Book, The MIT press, 1995;276-278.
- Ma Y., Xie Y. Evolutionary neural networks for deep learning: a review. *Int. J. Mach. Learn. Cyber.* 2022;13:3001-3018. DOI 10.1007/s13042-022-01578-8.
- Macquart D., Garnier L., Combe M., Beugnon G. Ant navigation *en route* to the goal: signature routes facilitate way-finding of *Gigantiops destructor*. *J. Comp. Physiol. A. Neuroethol. Sens. Neural. Behav. Physiol.* 2006;192(3):221-234. DOI 10.1007/s00359-005-0064-7.
- Moffett M.W., Garnier S., Eisenhardt K.M., Furr N.R., Warglien M., Sartoris C., Ocasio W., Knudsen T., Bach L.A., Offenberg J. Ant colonies: building complex organizations with minuscule brains and no leaders. *J. Org. Design.* 2021;10:55-74. DOI 10.1007/s41469-021-00093-4.
- Penick C.A., Ghaninia M., Haight K.L., Opachaloemphan C., Yan H., Reinberg D., Liebig J. Reversible plasticity in brain size, behavior and physiology characterizes caste transitions in a socially flexible ant (*Harpegnathos saltator*). *Proc. R. Soc. B. Biol. Sci.* 2021;288(1948):20210141. DOI 10.1098/rspb.2021.0141.
- Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington DC: Spartan Books, 1962.
- Sarma G.P., Lee C.W., Portegys T., Ghayoomie V., Jacobs T., Alicea B., Cantarelli M., Currie M., Gerkin R.C., Gingell S., Gleeson P., Gordon R., Hasani R.M., Idili G., Khayrulin S., Lung D., Palyanov A., Watts M., Larson S.D. OpenWorm: overview and recent advances in integrative biological simulation of *Caenorhabditis elegans*. *Philos. Trans. R. Soc. B. Biol. Sci.* 2018;373(1758):20170382. DOI 10.1098/rstb.2017.0382.
- Scheffer L.K., Xu C.S., Januszewski M., Lu Z., Takemura S.Y., Hayworth K.J., Huang G.B., ... Meinertzhagen I.A., Rubin G.M., Hess H.F., Jain V., Plaza S.M. A connectome and analysis of the adult *Drosophila* central brain. *eLife.* 2020;9:e57443. DOI 10.7554/eLife.57443.
- Stanley K.O., Miikkulainen R. Evolving neural networks through augmenting topologies. *Evol. Comput.* 2002;10(2):99-127. DOI 10.1162/106365602320169811.
- Stimberg M., Goodman D.F.M., Nowotny T. Brian2GeNN: accelerating spiking neural network simulations with graphics hardware. *Sci. Rep.* 2020;10(1):410. DOI 10.1038/s41598-019-54957-7.

ORCID ID

A. Palyanov orcid.org/0000-0003-1108-1486

Acknowledgements. The study was performed according to the Russian Federation Government research assignment for A.P. Ershov Institute of Informatics Systems SB RAS, project FWNU-2022-0006.

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

Received August 19, 2022. Revised November 1, 2022. Accepted November 8, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>


Human phospholipases A2: a functional and evolutionary analysis

I.I. Turnaev^{1, 3} , M.E. Bocharnikova^{2, 3}, 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 Novosibirsk State University, Novosibirsk, Russia

³ Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

 turn@bionet.nsc.ru

Abstract. Phospholipases A2 (PLA2) are capable of hydrolyzing the *sn*-2 position of glycerophospholipids to release fatty acids and lysophospholipids. The PLA2 superfamily enzymes are widespread and present in most mammalian cells and tissues, regulating metabolism, remodeling the membrane and maintaining its homeostasis, producing lipid mediators and activating inflammatory reactions, so disruption of PLA2-regulated lipid metabolism often leads to various diseases. In this study, 29 PLA2 genes in the human genome were systematically collected and described based on literature and sequence analyses. Localization of the PLA2 genes in human genome showed they are placed on 12 human chromosomes, some of them forming clusters. Their RVI scores estimating gene tolerance to the mutations that accumulate in the human population demonstrated that the G4-type PLA2 genes belonging to one of the two largest clusters (4 genes) were most tolerant. On the contrary, the genes encoding G6-type PLA2s (*G6B*, *G6F*, *G6C*, *G6A*) localized outside the clusters had a reduced tolerance to mutations. Analysis of the association between PLA2 genes and human diseases found in the literature showed 24 such genes were associated with 119 diseases belonging to 18 groups, so in total 229 disease/PLA2 gene relationships were described to reveal that G4, G2 and G7-type PLA2 proteins were involved in the largest number of diseases if compared to other PLA2 types. Three groups of diseases turned out to be associated with the greatest number of PLA2 types: neoplasms, circulatory and endocrine system diseases. Phylogenetic analysis showed that a common origin can be established only for secretory PLA2s (G1, G2, G3, G5, G10 and G12). The remaining PLA2 types (G4, G6, G7, G8, G15 and G16) could be considered evolutionarily independent. Our study has found that the genes most tolerant to PLA2 mutations in humans (G4, G2, and G7 types) belong to the largest number of disease groups.

Key words: phospholipase A2; glycerophospholipids; human diseases.

For citation: Turnaev I.I., Bocharnikova M.E., Afonnikov D.A. Human phospholipases A2: a functional and evolutionary analysis. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):787-797. DOI 10.18699/VJGB-22-95


Фосфолипазы A2 человека: функциональный и эволюционный анализ

И.И. Турнаев^{1, 3} , М.Е. Бочарникова^{2, 3}, Д.А. Афонников^{1, 2, 3}

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

³ Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

 turn@bionet.nsc.ru

Аннотация. Фосфолипазы A2 (PLA2) способны гидролизовать *sn*-2 положение глицерофосфолипидов для высвобождения жирных кислот и лизофосфолипидов. Ферменты семейства фосфолипазы A2 широко распространены и присутствуют в большинстве клеток и тканей млекопитающих, выполняя функции регулятора метаболизма, поддержания мембранного гомеостаза, производства липидных медиаторов, ремоделирования мембран, активации воспалительных реакций. Соответственно, нарушение PLA2-регулируемого липидного метаболизма часто приводит к различным заболеваниям. В настоящем исследовании были систематически собраны и описаны 29 генов PLA2 в геноме человека на основе анализа литературных данных и изучения последовательностей. Анализ локализации генов PLA2 в геноме человека показал, что они расположены на 12 хромосомах человека и некоторые из них образуют кластеры. Оценка значений величины RVI (оценка толерантности генов к мутациям, которые накапливаются в популяции человека) демонстрирует, что гены фосфолипаз A2 типа G4, входящие в один из двух наиболее крупных кластеров (четыре гена), наиболее толерантны к мутациям. Напротив, пониженную толерантность к мутациям имеют локализованные вне кластеров гены, кодирующие фосфолипазы A2 типа G6 (фосфолипазы A2 *G6B*, *G6F*, *G6C*, *G6A*). Мы проанализировали также связи между фосфолипаза-

ми A2 и заболеваниями человека по литературным данным, в результате чего выявлены связи 24 генов PLA2 со 119 заболеваниями, относящимися к 18 группам. Описано 229 связей «болезнь–ген» фосфолипазы A2. Показано, что белки фосфолипаз A2 типов G4, G2 и G7 вовлечены в наибольшее число заболеваний по сравнению с другими типами PLA2. С наибольшим числом типов PLA2 были связаны три группы заболеваний: новообразования, болезни системы кровообращения и болезни эндокринной системы. Филогенетический анализ показал, что общее происхождение устанавливается только для секреторных PLA2 (G1, G2, G3, G5, G10 и G12). Остальные типы PLA2 (G4, G6, G7, G8, G15 и G16) можно считать эволюционно независимыми. В результате проведенного анализа установлено, что наиболее толерантные к мутациям фосфолипазы A2 у человека (типы G4, G2 и G7) вовлечены в наибольшее количество групп заболеваний.

Ключевые слова: фосфолипаза A2; глицерофосфолипиды; заболевания человека.

Introduction

Phospholipases (PLs, EC 3.1) are hydrolases, enzymes that use a water molecule to degrade phospholipids (Burke, Dennis, 2009; Aloulou et al., 2018), the main component of the biological membranes of all living organisms (De Maria et al., 2007). There exist four classes of PLs (A, B, C, D), each of them being able to hydrolyze a specific bond in a phospholipid, e. g., phospholipase A1 (PLA1, EC 3.1.1.32) and phospholipase A2 (PLA2, EC 3.1.1.4) are acyl esterases and hydrolyze the *sn*-1 and *sn*-2 positions of glycerophospholipids, respectively; phospholipase B (PLB, EC 3.1.1.5) hydrolyzes both *sn*-1 and *sn*-2 positions of glycerophospholipids; phospholipases C (PLC, EC 3.1.4.3) and D (PLD, EC 3.1.4.4) are phosphate esterases and are determined based on the hydrolysis of glycerol or the distal side of the phosphate group (Fig. 1) (Aloulou et al., 2018; Shayman, Tesmer, 2019).

The PLA2 family is one being most extensively studied, which reflects their biological importance. They hydrolyze the ester bond of membrane phospholipids from the *sn*-2 position, and, under natural conditions, their *sn*-2 positions often contain polyunsaturated fatty acids, which, when released, can be metabolized to form various eicosanoids and their associated biologically active lipid mediators (Aloulou et al., 2018).

At least sixteen PLA2 types are known to the date. Dennis et al. (2011) divided them into six groups based on their properties: secreted phospholipases A2 (sPLA2, types G1, G2, G3, G5, G9, G10, G11, G12, G13 and G14); cytosolic phospholipases A2 (cPLA2, type G4); calcium-independent

phospholipases A2 (iPLA2, type G6); plasma platelet-activating factor acetylhydrolase (PAF-AH, types G7 and G8); lysosomal phospholipase A2 (LPLA2, type G15), and adipocyte phospholipase A2 (AdPLA, type G16).

Assigning a PLA2 to a certain group (type) is based on the experimental determination of their catalytic mechanisms, cellular localization, evolutionary and structural features. Note that most of these lipolytic enzymes share no structural similarity and have different regulatory and catalytic mechanisms (Aloulou et al., 2018).

Each of the sixteen PLA2 types is involved in lipid metabolism and disease development mechanisms of different kind, so PLA2s are believed to be promising therapeutic targets for a number of diseases (Aloulou et al., 2018). In this respect, there is a huge interest in the pharmaceutical industry for development of selective and effective inhibitors for each of these PLA2 types (Aloulou et al., 2018).

Describing protein functions is known to include, on the one hand, the molecular function, and, on the other, the function at the level of the vital activity of a cell or a whole organism (Karp, 2000). PLA2s have been fairly well studied in terms of their molecular functioning, however, their role in the vital processes of a cell and a whole organism remains poorly understood.

The objective of the present study was to analyze the characteristics of various human PLA2 types in the context of the available data on their association with various diseases. To do so, the PLA2s' protein-sequence domain organization, gene distribution in the genome, mutability characteristics as well as their phylogenetic relationships with the PLs of other organisms were analyzed.

Materials and methods

Sampling of human and animal PLs. The human PLA2-protein sequences were taken from Huang Q. et al. (2015), and since not all known human PLA2s were described in this paper (Dennis et al., 2011), the missing sequences were identified in the NCBI database by their names and identifiers as per Dennis et al. (2011), using the GRCh38.p14 human genome assembly.

The genome contained 29 PLA2 genes encoding twelve types of proteins (PLA2G1–8, 10, 12, 15, 16) (for sequences, see Suppl. Material 2)¹. PL types A2 G1, G3, G5, G10, G15,

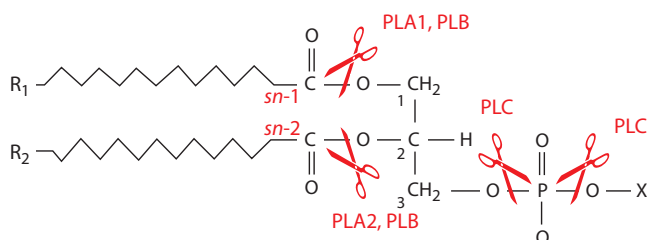


Fig. 1. Structural diagram of a phospholipid and the positions of the ester bonds hydrolyzed by different PL classes.

R₁ and R₂ are ((CH₂)_n · CH₃); X is various polar tail glycerophospholipid groups such as serine, choline, ethanolamine, glycerol or inositol; *sn*-1 and *sn*-2 are glycerophospholipid positions. Adapted from (Giresha, 2021).

¹ Supplementary Materials 1–11 are available in the online version of the paper: http://vavilov.elpub.ru/jour/manager/files/Suppl_Turnaev_Engl_26_8.pdf

G16 were represented by one gene; types G7 (*G7A*, *G7B*), G8 (*G8A*, *G8B*) and G12 (*G12A*, *G12B*) – by two genes; G2 – by five genes (*G2A*, *G2C–F*) and the G6 type was represented by six genes (*G6A–F*).

The primary structures of human PLA2s were characterized by the presence of domains, active sites, and signal peptides using the published data. To search for the PLA2s' homologues in animals, the BLASTP program (E-value ≤ 1) was employed with the human PLA2-protein sequences used as a query. The homologues were searched for among the protein sequences of the organisms representing various taxa, for their list see Suppl. Material 1.

Functional analysis of the PLA2s. To estimate the degree of the PLA2 genes' evolutionary conservation, the Residual Variance Index Score (RVIS) (Petrovski et al., 2013) was applied. The scoring enables one to assess a gene's tolerance to the mutations that accumulate in the human population, so the score is calculated based on the allele frequency information presented in the entire human exome sequence (data set NHLBI-ESP6500 from EVS v.0.0.14: <https://evs.gs.washington.edu/EVS/>). The score allows ranking genes by the number of observed nucleotide variations, taking into account the relative proportion of neutral substitutions that are observed for a gene under study. If negative, its value indicates low gene variability (i.e., its sequence is less tolerant to the accumulated mutations found in genes with a more important function), and if positive, it shows a higher gene variability (i.e., its sequence is more tolerant to nucleotide substitutions).

The DAVID service (Huang D.W. et al., 2009) was employed to identify the biological processes involving PLA2s. The service allows one to identify the terms from the Gene Ontology, INTERPRO and KEGG Pathway databases, over-represented in the annotations of the genes from an analyzed sample in comparison with the annotations of all genes in a human body. In our case, such a sample was a sample of human PLA2 genes.

Searching for PLA2/disease associations. The search for the articles describing the relationship between human diseases and PLA2-protein activity was carried out in the PubMed and Google Scholar databases using such queries as “disease/patients/pathology/name of a specific disease (e.g., lung cancer or schizophrenia) + PLA2/phospholipase A2/name of a specific PLA2 (e.g., pla2g1b, pla2g2a)”. Information was also taken from the reviews on PLA2 involvement in various diseases.

The found articles tracked information about the association of a person's disease and the activity/expression of a specific PLA2. For example, such information included reports about the patients who had significantly reduced/increased expression or activity of a certain PLA2 compared to healthy people; data that PLA2 gene mutation enhanced/weakened the severity of a disease; data that the mechanism enabling a PLA2 to influence the course of the disease had been established. To classify diseases in this study, the International Classification of Diseases (ICD-10 available at

<https://icd.who.int/browse10/2019/en>; in Russian at <https://mkb-10.com>) (Hirsch et al., 2016) was used.

Based on the information about the relationship between a disease and PLA2 involvement in it, a data table was formed, whose rows listed human disease types, and the columns – PLA2 types. If the table's cell had a value of 1, it meant this PLA2 type was involved/associated with the disease. To build this table, a Python script had been written, linking the name of a disease to its ICD-10 code.

As the next step, a hierarchical clustering of the human PLA2 types was performed according to the degree they were associated with various diseases. To do so, for different types of phospholipases, the degrees of their participation in the diseases from the abovementioned table were compared for different PLA2 types, using the Euclidean distance as a measure of similarity and the unweighted pair group method with arithmetic mean (UPGMA) – for clustering. In the same way, the diseases were clustered based on the degree of their association with different PLA2 types.

Multiple sequence alignment and protein phylogeny reconstruction. Multiple alignment of homologous PLA2 sequences was performed using the PROMALS (Pei, Grishin, 2007) and MAFFT (Kato, Toh, 2010) software. The search for proteins for alignment and, accordingly, the alignment of protein sequences were carried out only in the PL domain. The phylogenetic tree was reconstructed using the maximum likelihood method and the IQ-TREE software (v.8.2.4, see (Nguyen, 2015)) with an optimal WAG + R6 model chosen.

Results

Structural and functional characteristics of the human PLA2s

The features of the structural organization of the various types of the human PLA2s are shown in Figure 2. The proteins' properties (substrates, activity, mass, catalytic residues, etc.) is given in Suppl. Material 3.

Secretory phospholipases A2 (sPLA2). The sPLA2s included six types of PLA2s: G1, G2 (pla2g2(a, c–f)), G3, G5, G10, G12 (pla2g12(a, b)). The length of G1, G2, G5, G10 proteins was 138–165 aa, and that of G12 type – 189–195 aa. The G3-type protein was much larger and its size comprised 509 aa (see Fig. 2), which was due to the elongated C- and N-terminal extensions.

Phospholipids served as substrates for sPLA2 enzymes. In all cases, these were either phosphatidylcholine (PC) or phosphatidylethanolamine (PE), except for pla2g12a, whose substrate was phosphatidylglycerol (PG) but not PC or PE. Some sPLA2s also had PG and phosphatidylserine (PS) as substrates. The human pla2g12b protein was catalytically inactive (see Suppl. Material 3 and caption to Fig. 2).

Cytosolic phospholipases A2 (cPLA2). The cPLA2s were represented by the G4 type of PLA2 that included six human proteins: pla2g4(a–f). The mass of cPLA2 proteins varied from 541 aa (pla2g4c) to 1012 aa (pla2g4b) (see

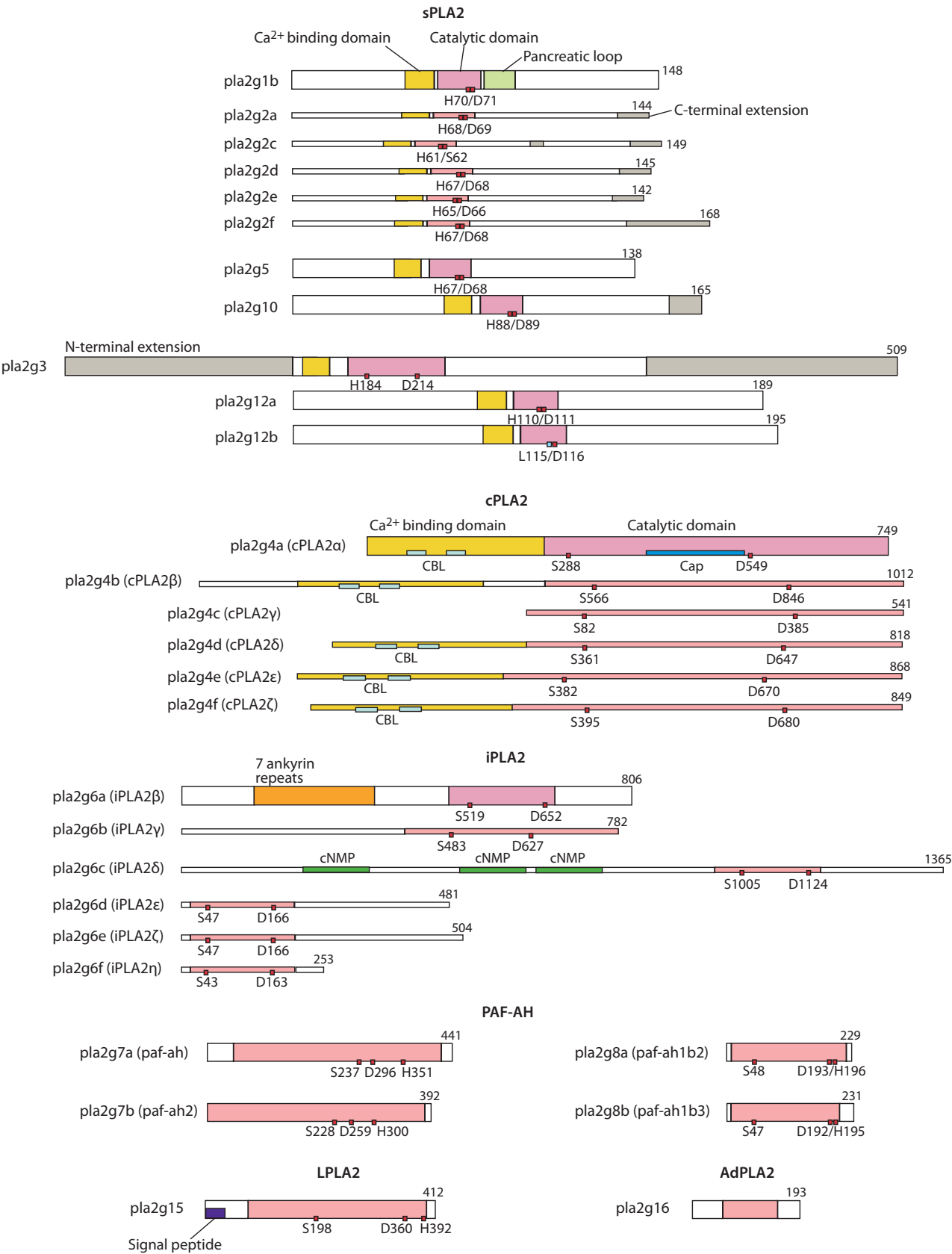


Fig. 2. Human PLA2 protein structure. The red rectangles mark the active sites, the blue one in the pla2g12b sequence denotes H (histidine) replaced by L (leucine) at position 115 of the protein that kills its catalytic activity (Guan et al., 2011). CBL is a Ca^{2+} binding loop. cNMP is a domain binding cyclic nucleotides (cAMP or cGMP). The pancreatic loop is sPLA2G1B of unique five-amino-acid extension. The Cap is a domain found in PLA2G4A that opens/closes an active site for PL substrate modeling. The drawing was adopted from (Kudo, Murakami, 2002; Dennis et al., 2011).

Fig. 2). In the proteins, the catalytic domains were located in the C-terminus of the sequences and contained a conservative Ser/Asp catalytic dyad (see Suppl. Material 3; Fig. 2). As sPLA2s, cPLA2s are calcium-dependent PLA2, so they also had a calcium binding domain closer to the N-terminus (Dennis et al., 2011).

In the G4 type proteins (cPLA2/PLA2G4), as well as in sPLA2 proteins, PLA2 activity was observed if their substrates were either PC or PE. The pla2g4a protein additionally had phosphatidylinositol (PI) as a substrate, while the pla2g4c protein had PC, but its specificity for PE was not demonstrated (see Suppl. Material 3).

Calcium-independent phospholipases A2 (iPLA2). The iPLA2s included only G6-type PLA2s, their length varying from 253 to 1365 aa. In iPLA2 proteins, the catalytic domains were located closer to the C-terminus in pla2g(a, c) and closer to the N-terminus in pla2g(d-f) (see Fig. 2). As that of cPLA2s, the protein's catalytic domain contained a conservative catalytic Ser/Asp dyad (see Suppl. Material 3; Fig. 2). As reflected in their name, iPLA2 catalytic activity was independent of Ca^{2+} presence, and, unlike the sPLA2s and cPLA2s, they did not have a Ca^{2+} -binding domain. The pla2g6a protein had a region containing 7 ankyrin repeats closer to the N-terminus. This motif is involved in protein-protein interactions, allowing intensive binding to membrane proteins (Filkin et al., 2020). In the pla2g6c protein, closer to the C-terminus were three cNMP sites (site-binding cyclic nucleotides) (see Fig. 2).

The g6(a-f) proteins used PC as substrates for PLA2 reactions; in the case of the pla2g6b protein, it could have PE in addition to PC. In addition to PLA2, these enzymes could also exhibit other activities such as TG-hydrolase, lysophospholipase, PLA1 (for pla2g6b) and other ones (see Suppl. Material 3).

Platelet activating factor acetyl hydrolase (PAF-AH or Lp-PLA2). The PAF-AHs included G7 and G8 PLA2s that modulated the activity of a platelet activating factor (PAF), a potent phospholipid inflammation mediator involved in inflammation, platelet aggregation and anaphylactic shock pathogenesis (Shimizu, 2009). The length of PAF-AH proteins was 441 and 392 aa for g7a and g7b, and 229 and 231 aa – for g8a and g8b, respectively (see Fig. 2). In the proteins, the catalytic domain occupied almost the entire sequence and contained a conserved Ser/His/Asp catalytic triad (see Suppl. Material 3; Fig. 2). They were independent of Ca^{2+} and had no Ca^{2+} binding domain (see Fig. 2).

Proteins pla2g7(a, b) and pla2g8(a, b) were able to hydrolyze a phospholipid platelet activating factor (PAF) into a lysoPAF. At the same time, the pla2g7a protein possessed both PLA2 and PLA1 activities and could use as a substrate both PC and oxPC. The pla2g7b protein showed a PLA2 activity (see Suppl. Material 3).

Lysosomal phospholipases A2 (LPLA2). The LPLA2s (G15-type PLA2s) were represented by a single pla2g15 protein of 412 aa in length (see Fig. 2). The protein's catalytic domain was located in the central region of the sequence

and contained a conservative Ser/His/Asp catalytic triad (see Suppl. Material 3; Fig. 2). The lpla2 protein was independent of Ca^{2+} and had no Ca^{2+} binding domain.

The pla2g15/lpla2 protein of type G15 had PLA2 and PLA1 activities, whose substrates being PC, PE and PS. Also, pla2g15 was capable of acylceramide synthase activity through C1 ceramide (see Suppl. Material 3).

Adipocyte phospholipases A2 (AdPLA2). The AdPLA2s (G16-type PLA2s) were represented by a single pla2g16 protein of 193 aa in length (see Fig. 2). In pla2g16, the catalytic domain was located in the central region of the sequence and, like in LPLA2s contained a conservative Ser/His/Asp catalytic triad (see Suppl. Material 3; Fig. 2). The adpla2 protein was independent of Ca^{2+} and, as iPLA2, PAF-AH, LPLA2, had no Ca^{2+} binding domain (see Fig. 2).

The protein had PLA2 and PLA1 activities through PC and PE substrates and a N-acyl-PE acyltransferase activity, through diacyl PE (see Suppl. Material 3).

PLA2-gene localization in human genome

The localization of the PLA2 genes in the human genome (version GRCh38.p14) is shown in Figure 3. The genes were absent on the 2, 3, 5, 8, 9, 13, 14, 17, 18, 20, 21st and Y chromosomes. The 4, 6, 7, 10, 12th and X chromosomes contained one PLA2 gene; chromosomes 11, 16 – two PLA2 genes; chromosomes 19, 22 – three PLA2 genes. In the 15th chromosome of the four genes (*G4B*, *G4E*, *G4D*, *G4F*) formed a 0.3 Mb cluster at the 43 Mb position. On chromosome 1, in addition to a single *G4A* gene (at 188 Mb), at the 20 Mb position was a cluster of six genes (*G2E*, *G2A*, *G5*, *G2D*, *G2F*, *G2C*) of 0.11 Mb in size. It is noteworthy that, excluding the genes of these two clusters, all other genes were isolated from one another at a distance of at least 6 Mb. Moreover, while the G4-type PLA2 genes (*G4B*, *G4E*, *G4D*, *G4F*) were located in the above-mentioned cluster on chromosome 15, the other two genes of this type (*G4A* and *G4C*) were isolated on chromosomes 1 and 19, respectively, so all PLA2 genes of type G2 (*G2A*, *G2C*, *G2D*, *G2E*, *G2F*) were located in a cluster on chromosome 1, but, together with them, this cluster included the *G5* gene. The G6-type PLA2 genes were located: *G6B* – on chromosome 7, *G6E* – on chromosome 11, *G6C* – on chromosome 19, *G6A* and *G6D* – on chromosome 22, and *G6F* – on the X chromosome.

RVIS-based PLA2-gene mutation tolerance

Figure 4 displays RVI-score distribution for human PLA2s. On the left of the graph are PLA2s whose score is above zero, so these are genes that contain a relatively large number of mutations and are tolerant to them. To the right are PLA2s whose score is below zero, so they are less tolerant to mutations. The genes of G16, G1, G12 (*PLA2G12B*), G4 (*PLA2G4A*), G5, G15, and G6 types had a negative RVI score (see Fig. 4). Of these PLA2s, three were secreted G1, G12 and G5 types as well as calcium-independent (type G6), cytosolic (G4), lysosomal (G15), and adipocyte (G16) PLA2 genes. Interestingly, four of the six PLA2 genes of G6 type

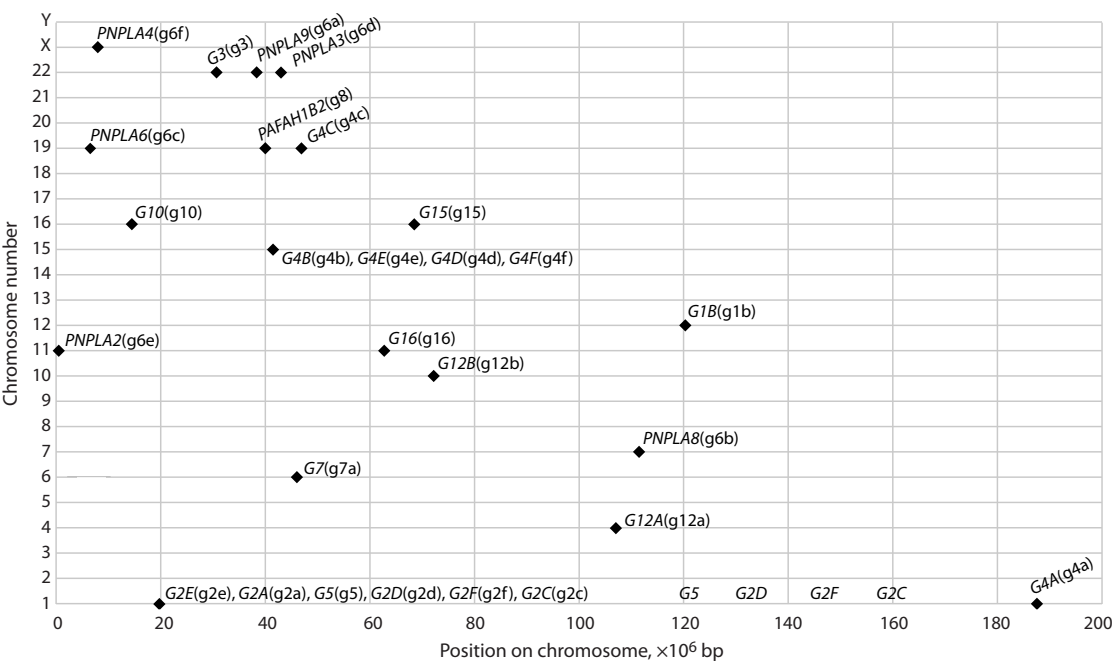


Fig. 3. PLA2-gene localization on human chromosomes.
The genes are marked as diamonds, their positions along the X axis correspond to the gene start coordinates on the chromosome, and along the Y axis – to the chromosome number.

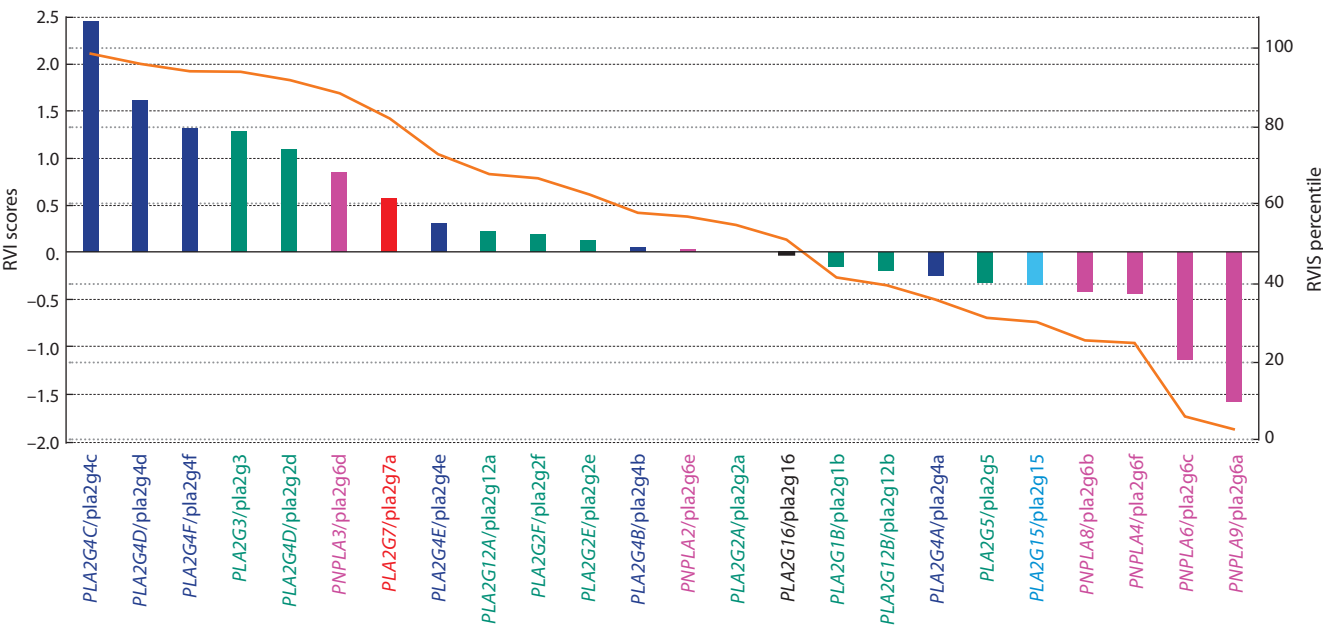


Fig. 4. RVI-score distribution for the human PLA2 genes.
The Y-axis's right part corresponds to the proportion of genes in the human genome (in %) whose RVI score is less than that for a particular gene (bar). These percentile values are marked on the graph as orange lines. The columns of other colors mark PLA2s of different types: cPLA2 (dark blue), sPLA2 (green), iPLA2 (pink), PAF-AH (red), adPLA2 (black), LPLA2 (blue).

(iPLA2s), in particular *PNPLA8*, *PNPLA4*, *PNPLA6* and *PLA2G6* (encoding proteins pla2g6b, pla2g6f, pla2g6c, pla2g6a) had the lowest RVI score, i.e., they were least tolerant to mutations, and the remaining genes (*PNPLA2* and *PNPLA3*, encoding proteins pla2g6e and pla2g6d) had the score indicating moderate or slightly increased tolerance.

The most tolerant to mutations were the cPLA2 genes of G4 type. Five out of the six genes of this kind had a positive RVI score and only one (*PLA2G4A*) – a score less than 0 (RVIS = –0.25). The most mutation-tolerant genes in this group (*PLA2G4(B, D–F)*) clustered on chromosome 15, unlike the *PLA2G4A* gene placed separately from them,

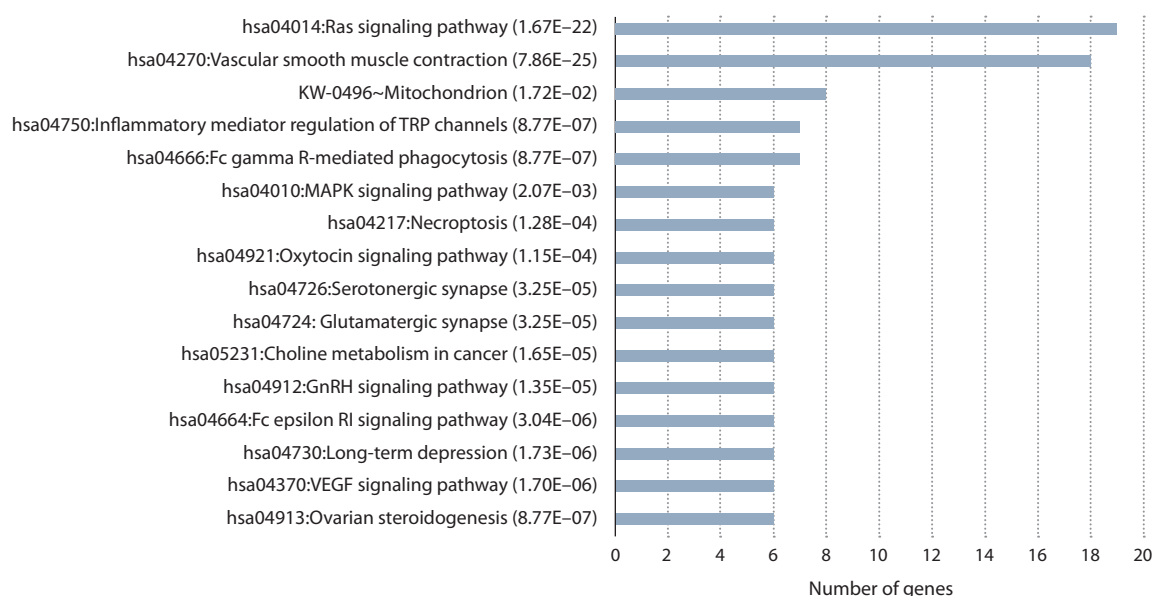


Fig. 5. Signaling pathways and biological processes from the KEGG Pathway database that were detected by the DAVID service as significantly associated with the found PLA2 genes.

Along the Y-axis are terms describing the signaling pathways and biological processes. In brackets, after each term, the false discovery rate value (FDR or expected proportion of false rejections) is given. The X-axis plots the number of PLA2 genes associated with each term.

on chromosome 1. For five human PLA2s (*PLA2G2C*, *PLA2G7B*, *PLA2G8A*, *PLA2G8B*, *PLA2G10*), the EVS server did not contain any gene-variability data to calculate their RVI score (see Materials and methods section), so they were excluded from the graph (see Fig. 4).

Human PLA2 relationship to the biological processes and signaling pathways from the KEGG Pathway database

Figure 5 demonstrates the results of a functional analysis of the found PLA2 genes performed using the DAVID service. It turned out that the most significant (based on the number of PLA2 genes associated with it) was the Ras signaling pathway that was involved in carcinogenesis. Another most used term was the VEGF pathway associated with a vascular endothelial growth factor. The diseases associated with this pathway also tended to be associated with the development of such tumors as breast cancer, glioma, melanoma, etc. (Takahashi, Shibuya, 2005). Thus, the data have shown that PLA2s are significantly associated with carcinogenesis.

PLA2-associated diseases

A relationship between the activity of various PLA2s and human diseases was analyzed based on published papers; the results are given in Suppl. Material 4. The diseases are designated according to the International Classification of Diseases (ICD-10, available at www.mkb-10.com). The table reflects the relationship between a disease and PLA2 in cases where the articles contain information on: (1) the association between the expression of a certain PLA2 and the course of the disease, or (2) the association between PLA2 mutations and the course of the disease, or (3) the mechanisms PLA2 affects the course of the disease. The table

describes 229 disease-gene associations and demonstrates the associations between the 24 PLA2 genes belonging to 12 PLA2 types, and 119 diseases (see Suppl. Material 4).

The PLA2s of various types were clustered based on their associations with human diseases. The results are shown in Figure 6. Eighteen disease groups, their names and ICD-10 codes (in parentheses) are the rows of the clustering diagram. The bars in the diagram correspond to the 12 types of human PLA2s. Most PLA2 groups were associated with neoplasms (ICD-10 code: C00–D48; 9 groups in 12); diseases of the circulatory system (I00–I99; 8 in 12); diseases of the endocrine system (E00–E90; 7 in 12); diseases of the eye and adnexa (H00–H59; 6 in 12). The smallest number of PLA2 groups was associated with congenital anomalies (Q00–Q99; only one G7-type PLA2); symptoms, signs and abnormalities (R00–R99; one G6-type PLA2); certain infectious and parasitic diseases (A00–B99; only G2 and G7-type PLA2s).

It is interesting to note that out of considered PLA2 types, the following had most associations with diseases: G7 was associated with 15 disease groups out of the 18 presented in Figure 6; G2 – with 13 groups; G4 – with 12 groups. The least represented in the disease groups were: G8 associated only with diseases of the genitourinary system (N00–N99); G15 – only with diseases of the circulatory system (I00–I99); G12 – only with mental and behavioral disorders (F00–F99) and with diseases of the eye and adnexa (H00–H59); G16 – only with neoplasms (C00–D48) and with diseases of the endocrine system (E00–E90).

The horizontal clustering of the PLA2s demonstrated their division into three clusters (see Fig. 6). The first contained the G4, G2, G7 types and the genes were involved in a large number of the human diseases analyzed. The second cluster

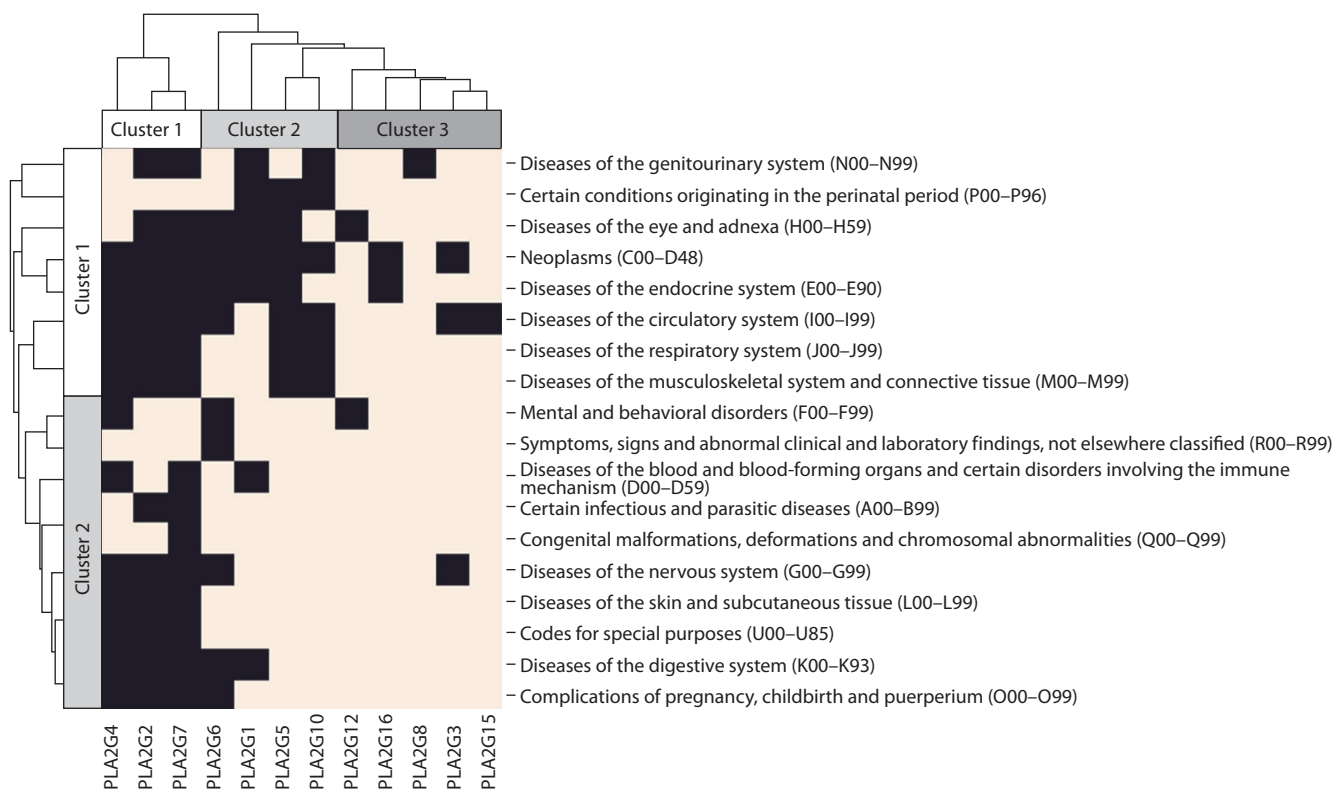


Fig. 6. PLA2 associations with various disease groups.

The cells are colored black when links between the genes of a PLA2 group and the diseases from the presented disease groups have been revealed. The white color marks the cases when no gene-disease links have been identified.

included the G6, G5, G10 types being secreted PLA2s. They are involved in about a half of the analyzed diseases including diseases of various systems: genitourinary (N00–N99), circulation (I00–I99), respiration (J00–J99); neoplasms (C00–D48), etc. The third cluster (G12, G16, G8, G3, G15) can be considered specific for individual diseases.

The diseases, on the other hand, can also be divided into two broad groups: those involving most PLA2 types (cluster 1) and the diseases (cluster 2) involving PLA2s mainly belonging to the first PLA2 cluster (see Fig. 6).

It is noteworthy that of the twelve studied PLA2 genes of G4, G2, G7 types, eleven had a high level of mutation tolerance (RVIS) and only one, *PLA2G4A*, had a moderately low level of mutation tolerance (RVIS = –0.25) (see Fig. 4), given that these types are involved in the greatest number of diseases (see Fig. 6). At the same time, of the seven studied PLA2 genes of types G6 and G15, five had the lowest level of tolerance to mutations (RVIS) and only one, *PLA2G6D*, had a relatively high level of tolerance to mutations (RVIS = 0.85) (see Fig. 4), given that these types belong to the cluster associated with the least number of diseases (see Fig. 6). This suggests a possible positive relationship between the number of diseases in which a PLA2 is involved and the gene's mutation tolerance (RVI score). However, calculating the correlation coefficient by the χ^2 method did not reveal a significant correlation between these values, so, in this case, we can only speak of an unreliable trend.

At the same time, Petrovski et al. (2013) studying a sample of the genes associated with Mendelian (monogenic) diseases, demonstrated that they had a low tolerance (RVI score) compared to other human genes. The authors suggested that a negative RVI score indicated the presence of purifying selection, and a positive one – either the absence of purifying selection, or even the presence of some form of balanced or positive selection.

The tendency towards an increased RVI score in the PLA2 genes involved in a greater number of diseases, may be due to the fact that, when considering expression data, a signal from a set of identified differentially-expressed genes can be significantly contaminated by the noisy produced by random genes. The appearance of such random genes can be associated as with the features of an applied technique (Hatfield et al., 2003) as with the fact that any perturbation in the cell and organism (e. g., a disease) can induce nonspecific effects on gene expression (e. g., stress response genes activation, apoptosis, necrosis, etc.) (Leuner et al., 2007; Turkmen, 2017). Therefore, when estimating the number of associations between a PLA2 and a disease, both large and small numbers of associations must be interpreted with some caution.

PLA2 evolution

Searching for homologous sequences in the protein databases was employed to identify PLA2 sequences for 32 species

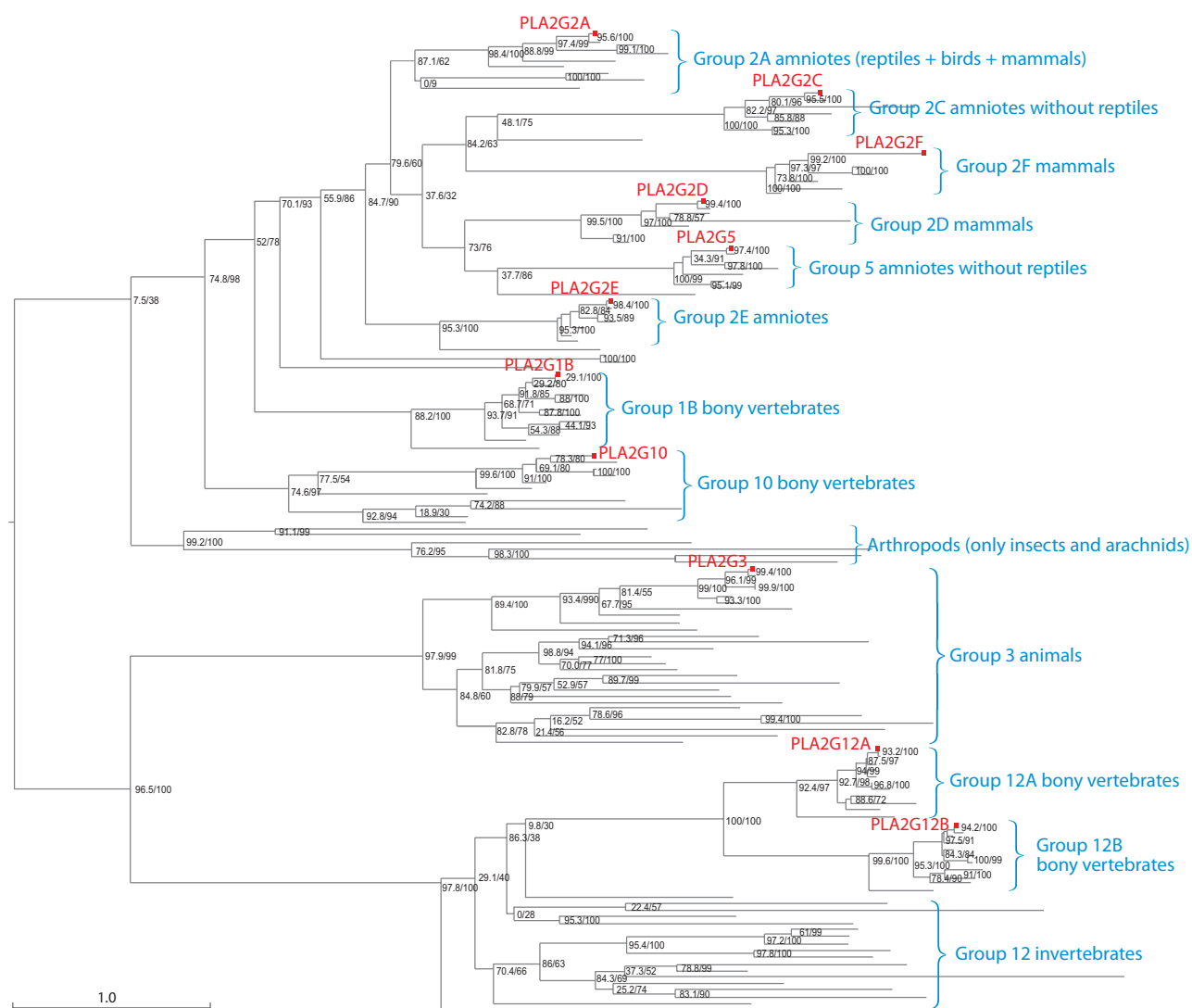


Fig. 7. PLA2 phylogenetic tree: G1, G2, G3, G5, G10, G12 types.

The type names (clusters on the tree) are given in blue text that describes which taxa are represented in each cluster. The red texts and squares highlight human PLA2 proteins. Their two types of bootstrap support are shown next to the tree nodes separated by a slash: ultrafast bootstrap (UFBoot)/bootstrap SH-aLRT. A textual description of the tree is given in Suppl. Material 11.

(see Suppl. Material 1), including 13 vertebrates and 19 invertebrates (see the Materials and methods section). Their identifiers and sequences are given in Suppl. Material 5.

To illustrate the similarity of PLA2 functional regions, a homology analysis of the catalytic domains of the human PLA2s was performed, and such a similarity between PLA2 domains of different types was only found among secretory PLA2s (Suppl. Material 6), in particular between the catalytic domains of G1, G2, G5, G10 types the E-value varied from $2e-03$ to $2e-38$; and between G12- type proteins (g12a and g12b) it was equal to $4e-48$. At the same time, no similarity (E-value ≥ 1) was detected between G3-type PLA2s (plag3) and all other sPLA2 proteins.

For all other types of PLA2, except for those belonging to sPLA2s, no similarity of domains between PLA2s of different types was found. Within types, in particular, between cPLA2s (type G4) the E-value varied from $4e-27$ to

$1e-177$ (Suppl. Material 7). Whereas iPLA2s (type G6) fell into three subgroups in this respect: (1) pla2g6d, 6e, 6f with the E-value varying from $2e-42$ to $4e-91$; (2) pla2g6a, 6b whose similarity between their catalytic domains was $4e-12$; (3) the pla2g6c catalytic domain that had no homological similarity with any other type-6 proteins (Suppl. Material 8). Respectively, there was no similarity between the human PLA2 domains of these three G6 subtypes. G7 (two proteins g7a and g7b) and G8 (two proteins g8a and g8b) types had similarities within these type of sequences: $7e-103$ (G7) (Suppl. Material 9) and $3e-102$ (G8) (Suppl. Material 10). For the remaining two types, no comparison was made, since in human, they included only one protein each.

To reconstruct the phylogeny of the PLA2 proteins, a homology and multiple sequence alignment analysis for the proteins of different PLA2 types had been initially performed. It showed that the proteins of the secreted sPLA2

group (G1–3, G5, G10, G12 types) had high or moderate homology (E -value ≤ 1) and qualitative alignment within the group. In contrast, the proteins of other PLA₂ types (G4, G6, G7, G8, G15, G16) had very low homology (E -value > 1) and poorly aligned as between themselves as with respect to the sPLA₂ proteins. In this respect, sPLA₂ phylogeny was reconstructed using the maximum likelihood method (Fig. 7).

The results of the phylogenetic analysis enabled us to assume that two successive divergences occurred in the common ancestors of multicellular invertebrates: first, the ancestral sPLA₂ gene diverged into the *G3/G12* and *G1/G2/G5/G10* ones, and then the *G3/G12* gene diverged into the ancestral *G3* and *G12* genes.

In the common ancestors of bony vertebrates, the ancestral *G12* gene diverged into the *G12A* and *G12B* genes, while in the common ancestors of bony vertebrates, the ancestral *G1/G2/G5/G10* gene diverged into the *G10* and *G1/G2/G5* genes, and then the *G1/G2/G5* gene diverged into the *G1* and *G2/G5* genes. Further, in the common ancestors of amniotes, the *G2/G5* gene diverged into the *G2E* and *G2A/G2C/G2D/G2F/G5* genes, then the *G2A/G2C/G2D/G2F/G5* gene – into the *G2A* and *G2C/G2D/G2F/G5* genes, then the gene *G2C/G2D/G2F/G5* – to the *G2C/G2F* and *G2D/G5* genes, and as a result, the *G2C/G2F* gene diverged into the *G2C* and *G2F* genes, and the *G2D/G5* gene – into the *G2D* and *G5* genes. Thus, G2-type PLA₂s appear to be paraphyletic, as it also includes a cluster of G5-type PLA₂s, whereas all the other sPLA₂ groups are monophyletic.

Conclusion

The paper presents the results of analysis of the PLA₂ family in human and describes the structure and functions of 29 PLA₂s belonging to 12 types: G1–8, G10, G12, G15, G16. Analysis of PLA₂-gene localizations in the human genome has demonstrated they present on 12 chromosomes and some of them form clusters, the two largest of them include, first, all G2-type PLA₂ genes (*G2A*, *G2C–F*) and the *G5* gene, and second – G4 type PLA₂ genes (*G4B*, *G4D–F*).

The association between the PLA₂s and human diseases as they described in the literature have also been analyzed. In total, 229 disease–PLA₂ gene links have been found, so associations between 24 PLA₂ genes and 119 diseases have been demonstrated. The PLA₂ proteins of types G4, G2 and G7 have turned out to be involved in the greatest number of diseases if compared to the other types, whereas three groups of diseases have turned out to be associated with the largest number of PLA₂ types: neoplasms, circulatory- and endocrine-system diseases.

RVI scoring of the genes' tolerance/intolerance mutations has showed that the majority of genes of the G4 (*G4B*, *G4C*, *G4D*, *G4E*, *G4F*) and G2 (*G4D*, *G4E*, *G4F*) types, as well as the genes of the types represented by one G3 and G7 gene, were tolerant to mutations, whereas most genes of the G6 type (*G6A–C*, *G6F*) as well as the types represented by a single gene (G5 and G15), turned out to be

not tolerant. Here it should be noted that all the PLA₂ types with predominance of genes tolerant to mutations, except for G3, have also been associated with the greatest number of diseases: G4 (12 disease groups), G2 (13), G7 (15), while all the PLA₂ types intolerant to mutations have been associated with a smaller number of disease groups: G6 (9 disease groups), G6 (7), G6 (1), which suggests that higher tolerance to mutations in a particular human PLA₂ gene is associated with its involvement in more diseases or disease groups.

Phylogenetic analysis has demonstrated that a common origin can only be established for sPLA₂s (G1, G2, G3, G5, G10, G12), while the other investigated types (G4, G6, G7, G8, G15, G16) can be considered evolutionarily independent.

References

- Aloulou A., Rahier R., Arhab Y., Noiriell A., Abousalham A. Phospholipases: an overview. *Methods Mol. Biol.* 2018;1835:69-105. DOI 10.1007/978-1-4939-8672-9_3.
- Burke J.E., Dennis D.A. Phospholipase A₂ biochemistry. *Cardiovasc. Drugs Ther.* 2009;23(1):49-59. DOI 10.1007/s10557-008-6132-9.
- De Maria L., Vind J., Oxenbøll K.M., Svendsen A., Patkar S. Phospholipases and their industrial applications. *Appl. Microbiol. Biotechnol.* 2007;74(2):290-300. DOI 10.1007/s00253-006-0775-x.
- Dennis E.A., Cao J., Hsu Y.-H., Magriotti V., Kokotos G. Phospholipase A₂ enzymes: physical structure, biological function, disease implication, chemical inhibition, and therapeutic intervention. *Chem. Rev.* 2011;111(10):6130-6185. DOI 10.1021/cr200085w.
- Filkin S.Yu., Lipkin A.V., Fedorov A.N. Phospholipase superfamily: structure, functions, and biotechnological applications. *Uspekhi Biologicheskoi Khimii = Biochemistry (Moscow)*. 2020;85(Suppl.1): S177-S195. DOI 10.1134/S0006297920140096.
- Giresha A.S. Secretory phospholipase A₂ group IIA: a potential therapeutic target in inflammation. In: Kumar D. (Ed.) *Current Research and Trends in Medical Science and Technology*. Lucknow (Uttar Pradesh, India): Department of Ortho KGMU, 2021;1:34-85.
- Guan M., Qu L., Tan W., Chen L., Wong C.-W. Hepatocyte nuclear factor-4 alpha regulates liver triglyceride metabolism in part through secreted phospholipase A₂ GXIIB. *Hepatology*. 2011;53(2):458-466. DOI 10.1002/hep.24066.
- Hatfield G.W., Hung S.-P., Baldi P. Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.* 2003;47(4):871-877. DOI 10.1046/j.1365-2958.2003.03298.x.
- Hirsch J.A., Nicola G., McGinty G., Liu R.W., Barr R.M., Chittle M.D., Manchikanti L. ICD-10: history and context. *Am. J. Neuroradiol.* 2016;37(4):596-599. DOI 10.3174/ajnr.A4696.
- Huang D.W., Sherman B.T., Zheng X., Yang J., Imamichi T., Stephens R., Lempicki R.A. Extracting biological meaning from large gene lists with DAVID. *Curr. Protoc. Bioinformatics*. 2009;27: 13.11.1-13.11.13. DOI 10.1002/0471250953.bi1311s27.
- Huang Q., Wu Y., Qin C., He W., Wei X. Phylogenetic and structural analysis of the phospholipase A₂ gene family in vertebrates. *Int. J. Mol. Med.* 2015;35(3):587-596. DOI 10.3892/ijmm.2014.2047.
- Karp P.D. An ontology for biological function based on molecular interactions. *Bioinformatics*. 2000;16(3):269-285. DOI 10.1093/bioinformatics/16.3.269.
- Katoh K., Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*. 2010;26(15):1899-1900. DOI 10.1093/bioinformatics/btq224.
- Kudo I., Murakami M. Phospholipase A₂ enzymes. *Prostaglandins Other Lipid Mediat.* 2002;68-69:3-58. DOI 10.1016/s0090-6980(02)00020-5.

- Leuner K., Pantel J., Frey C., Schindowski K., Schulz K., Wegat T., Maurer K., Eckert A., Müller W.E. Enhanced apoptosis, oxidative stress and mitochondrial dysfunction in lymphocytes as potential biomarkers for Alzheimer's disease. *J. Neural. Transm. Suppl.* 2007; 72:207-215. DOI 10.1007/978-3-211-73574-9_27.
- Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 2015;32(1):268-274. DOI 10.1093/molbev/msu300.
- Pei J., Grishin N.V. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics.* 2007; 23(7):802-808. DOI 10.1093/bioinformatics/btm017.
- Petrovski S., Wang Q., Heinzen E.L., Allen A.S., Goldstein D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* 2013;9(8):e1003709. DOI 10.1371/journal.pgen.1003709.
- Shayman J.A., Tesmer J.J.G. Lysosomal phospholipase A₂. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids.* 2019;1864(6):932-940. DOI 10.1016/j.bbalip.2018.07.012.
- Shimizu T. Lipid mediators in health and disease: enzymes and receptors as therapeutic targets for the regulation of immunity and inflammation. *Annu. Rev. Pharmacol. Toxicol.* 2009;49:123-150. DOI 10.1146/annurev.pharmtox.011008.145616.
- Takahashi H., Shibuya M. The vascular endothelial growth factor (VEGF)/VEGF receptor system and its role under physiological and pathological conditions. *Clin. Sci.* 2005;109(3):227-241. DOI 10.1042/CS20040370.
- Turkmen K. Inflammation, oxidative stress, apoptosis, and autophagy in diabetes mellitus and diabetic kidney disease: the Four Horsemen of the Apocalypse. *Int. Urol. Nephrol.* 2017;49(5):837-844. DOI 10.1007/s11255-016-1488-4.

ORCID ID

I.I. Turnaev orcid.org/0000-0002-0448-1468
D.A. Afonnikov orcid.org/0000-0001-9738-1409

Acknowledgements. The study was supported by budget project No. FWNR-2022-0020. Data processing was carried out using the computing resources of the "Bioinformatics" Joint Computational Center ICG SB RAS and Novosibirsk State University Supercomputer Center.

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

Received September 11, 2022. Revised November 24, 2022. Accepted November 25, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>

Promoters of genes encoding β -amylase, albumin and globulin in food plants have weaker affinity for TATA-binding protein as compared to non-food plants: *in silico* analysis

O.V. Vishnevsky¹, I.V. Chadaeva¹, E.B. Sharypova¹, B.M. Khandaev¹, K.A. Zolotareva¹, A.V. Kazachek¹, P.M. Ponomarenko¹, N.L. Podkolodny^{1,2}, D.A. Rasskazov¹, A.G. Bogomolov¹, O.A. Podkolodnaya¹, L.K. Savinkova¹, E.V. Zemlyanskaya¹, M.P. Ponomarenko¹ ✉

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pon@bionet.nsc.ru

Abstract. It is generally accepted that during the domestication of food plants, selection was focused on their productivity, the ease of their technological processing into food, and resistance to pathogens and environmental stressors. Besides, the palatability of plant foods and their health benefits could also be subjected to selection by humans in the past. Nonetheless, it is unclear whether in antiquity, aside from positive selection for beneficial properties of plants, humans simultaneously selected against such detrimental properties as allergenicity. This topic is becoming increasingly relevant as the allergization of the population grows, being a major challenge for modern medicine. That is why intensive research by breeders is already underway for creating hypoallergenic forms of food plants. Accordingly, in this paper, albumin, globulin, and β -amylase of common wheat *Triticum aestivum* L. (1753) are analyzed, which have been identified earlier as targets for attacks by human class E immunoglobulins. At the genomic level, we wanted to find signs of past negative selection against the allergenicity of these three proteins (albumin, globulin, and β -amylase) during the domestication of ancestral forms of modern food plants. We focused the search on the TATA-binding protein (TBP)-binding site because it is located within a narrow region (between positions –70 and –20 relative to the corresponding transcription start sites), is the most conserved, necessary for primary transcription initiation, and is the best-studied regulatory genomic signal in eukaryotes. Our previous studies presented our publicly available Web service Plant_SNP_TATA_Z-tester, which makes it possible to estimate the equilibrium dissociation constant (K_D) of TBP complexes with plant proximal promoters (as output data) using 90 bp of their DNA sequences (as input data). In this work, by means of this bioinformatics tool, 363 gene promoter DNA sequences representing 43 plant species were analyzed. It was found that compared with non-food plants, food plants are characterized by significantly weaker affinity of TBP for proximal promoters of their genes homologous to the genes of common-wheat globulin, albumin, and β -amylase (food allergens) ($p < 0.01$, Fisher's Z-test). This evidence suggests that in the past humans carried out selective breeding to reduce the expression of food plant genes encoding these allergenic proteins.

Key words: food allergen; albumin; globulin; β -amylase; gene; promoter; common wheat *Triticum aestivum* L. (1753); plants; TATA-binding protein; TATA box; domestication; selection; *in silico* estimate.

For citation: Vishnevsky O.V., Chadaeva I.V., Sharypova E.B., Khandaev B.M., Zolotareva K.A., Kazachek A.V., Ponomarenko P.M., Podkolodny N.L., Rasskazov D.A., Bogomolov A.G., Podkolodnaya O.A., Savinkova L.K., Zemlyanskaya E.V., Ponomarenko M.P. Promoters of genes encoding β -amylase, albumin and globulin in food plants have weaker affinity for TATA-binding protein as compared to non-food plants: *in silico* analysis. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):798-805. DOI 10.18699/VJGB-22-96

Промоторы генов, кодирующих β -амилазу, альбумин и глобулин пищевых растений в сравнении с непищевыми, характеризуются более низкой аффинностью к ТАТА-связывающему белку: *in silico* анализ

О.В. Вишневецкий¹, И.В. Чадаева¹, Е.Б. Шарыпова¹, Б.М. Хандаев¹, К.А. Золотарева¹, А.В. Казачек¹, П.М. Пономаренко¹, Н.Л. Подколодный^{1,2}, Д.А. Рассказов¹, А.Г. Богомолов¹, О.А. Подколodная¹, Л.К. Савинкова¹, Е.В. Землянская¹, М.П. Пономаренко¹ ✉

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ pon@bionet.nsc.ru

© Vishnevsky O.V., Chadaeva I.V., Sharypova E.B., Khandaev B.M., Zolotareva K.A., Kazachek A.V., Ponomarenko P.M., Podkolodny N.L., Rasskazov D.A., Bogomolov A.G., Podkolodnaya O.A., Savinkova L.K., Zemlyanskaya E.V., Ponomarenko M.P., 2022

This work is licensed under a Creative Commons Attribution 4.0 License

Аннотация. Принято считать, что при доместикации пищевых растений отбор шел на урожайность, технологичность переработки в продукты питания, устойчивость к патогенам и стрессовым воздействиям окружающей среды. При этом также могли оцениваться вкусовые качества продуктов питания растительно-го происхождения и их ценность для здоровья. Однако неясно, проводил ли человек в прошлом наряду с положительным отбором на полезные свойства растений одновременно отбор против таких вредоносных свойств, как способность вызывать аллергические реакции. Этот вопрос становится все более актуальным по мере роста аллергизации населения как вызова современной медицине. В связи с этим селекционерами уже ведутся интенсивные исследования по созданию гипоаллергенных форм пищевых растений. В этой работе рассмотрены альбумин, глобулин и β -амилаза мягкой пшеницы *Triticum aestivum* L. (1753), идентифицированные ранее как мишени для атак иммуноглобулинов класса Е человека. Нашей целью было найти на геномном уровне следы отрицательного отбора в прошлом против гипераллергенности трех белков (альбумин, глобулин и β -амилаза) при одомашнивании предковых форм современных пищевых растений. Для этого мы сфокусировали поиск на сайте связывания ТАТА-связывающего белка (ТБП) как локализованном в узком районе $[-70; -20]$ относительно старта транскрипции, консервативном, необходимом для первичной инициации транскрипции и наиболее изученном регуляторном сигнале в геномах эукариот. Ранее нами был создан свободно доступный веб-сервис Plant_SNP_TATA_Z-tester для оценки величин равновесной константы диссоциации (K_D) комплексов ТБП с проксимальными промоторами генов растений по их последовательностям ДНК длиной 90 п. о. В настоящей работе с его помощью проанализированы 363 последовательности ДНК промоторов генов 43 видов растений. Обнаружено, что пищевые растения, в сравнении с непищевыми, характеризуются достоверно более низкой аффинностью ТБП к проксимальным промоторам их генов, гомологичных генам глобулина, альбумина и β -амилазы мягкой пшеницы как пищевых аллергенов ($p < 0.01$, Z-критерий Фишера). Это свидетельствует об отборе при доместикации пищевых растений в прошлом на снижение уровня данных аллергенных белков.

Ключевые слова: пищевые аллергены; альбумин; глобулин; β -амилаза; ген; промотор; мягкая пшеница *Triticum aestivum* L. (1753); растения; ТАТА-связывающий белок; ТАТА-бокс; доместикация; отбор; оценки *in silico*.

Introduction

Currently, the problem of food allergenicity is extremely relevant because the documented rapid growth of population allergization is becoming one of the key challenges for modern medicine (Prescott et al., 2022). In this regard, modern plant breeders are working in two directions: (1) creation of new hypoallergenic forms of agricultural food plants and (2) identification of new plant food allergens and of molecular mechanisms of their action (Hong et al., 2021; Cavazza et al., 2022).

The aim of our work was to search at the molecular genetic level for signs of negative selection against allergens during the domestication of ancestral forms of modern food plants. Three food allergens from common wheat *Triticum aestivum* L. (1753) were studied: β -amylase, albumin, and globulin, previously identified as targets of allergic reactions mediated by human class E immunoglobulins (Wang et al., 2021).

The current study was conducted using our previously developed freely available Web service Plant_SNP_TATA_Z-tester, which is designed to estimate the equilibrium dissociation constant (K_D) of a complex of *Arabidopsis thaliana* (L.) Heynh. (1842) TBP-1 (hereafter: "plant TBP") with a proximal promoter of various plant genes (Rasskazov et al., 2022). This tool was utilized to analyze 363 nucleotide sequences of proximal promoters of relevant genes from 43 plant species. As a result, compared to non-food plants, food plants were found to have significantly weaker affinity of plant TBP toward promoters of genes homologous to common-wheat genes of β -amylase, albumin, and globulin (food allergens). These data indicate that in the past, selection was carried out by humans for reducing the expression of food plant genes encoding allergenic proteins when such plants were domesticated.

Materials and methods

Nucleotide sequences of plant gene promoters analyzed in this work. Three allergenic proteins from common wheat *T. aestivum* were investigated: β -amylase, albumin, and globulin, which have previously been experimentally identified as targets for human class E immunoglobulins (Wang et al., 2021). From the GenBank database (Benson et al., 2015), nucleotide sequences of 90 bp proximal promoters were retrieved that are located immediately upstream of transcription start sites of plant genes homologous to the genes of β -amylase, albumin, and globulin from common wheat *T. aestivum*. After the exclusion of promoter DNA sequences with unknown nucleotides w, s, r, y, k, m, b, d, h, v, and n (according to the nomenclature of (IUPAC-IUB..., 1970)), we had 363 promoter sequences belonging to 43 plant species. Then, all 43 plant species were categorized into two nonoverlapping groups: group I, represented by 235 proximal promoters from 28 food plant species for which there was information about their centuries-old use by humans as foods (Table 1), and group II, represented by 128 proximal promoters from non-food plants (the other 15 species) (Table 2).

Nucleotide sequence analysis of proximal promoters of plants. Using Web service Plant_SNP_TATA_Z-tester (Rasskazov et al., 2022), which we have created earlier, we calculated K_D (in moles per liter; M) for complexes of plant TBP with each promoter by means of the nucleotide sequence of each promoter (characterized in Tables 1 and 2).

The calculations were performed in accordance with our previously formulated model of three-step binding of TBP to a promoter (i) TBP slides along the double helix of promoter DNA (Coleman, Pugh, 1995) \leftrightarrow (ii) TBP stops at a potential site of TBP binding (Berg, von Hippel, 1987; Bucher,

Table 1. Characteristics of 235 nucleotide sequences of proximal promoters of food plant genes homologous to the studied globulin (*Glo*), albumin (*Alb*), and β -amylase (*Bmy*) genes from common wheat *T. aestivum*

Food plant species		Number of promoters		
No.	Name	<i>Glo</i>	<i>Alb</i>	<i>Bmy</i>
1	Buckwheat <i>Fagopyrum esculentum</i> Moench, 1794	1	–	–
2	Maidenhair tree <i>Ginkgo biloba</i> L., 1771	1	–	–
3	Yoshino cherry <i>Prunus yedoensis</i> var. <i>nudiflora</i> Koehne, 1912	1	–	2
4	Maize <i>Zea mays</i> L., 1753	1	–	–
5	Oat <i>Avena sativa</i> L., 1753	2	–	–
6	Waxberry <i>Morella rubra</i> Siebold & Zucc.	2	2	1
7	Quinoa <i>Chenopodium quinoa</i> Willd., 1798	2	–	–
8	Rice <i>Oryza sativa</i> L., 1753	3	–	–
9	Melon <i>Cucumis melo</i> L., 1753	4	2	6
10	Cardoon <i>Cynara cardunculus</i> L.	4	–	–
11	Cork oak <i>Quercus suber</i> L.	4	–	–
12	Wine grape <i>Vitis vinifera</i> L.	9	–	9
13	Congolese coffee <i>Coffea canephora</i> Pierre ex A. Froehner, 1897	1	–	–
14	Pepper <i>Capsicum annuum</i> L., 1753	26	8	27
15	Sesame <i>Sesamum indicum</i> L.	–	1	–
16	Kiwifruit nashi-kazura <i>Actinidia rufa</i> Franch. & Sav.	–	1	–
17	Brazil nut <i>Bertholletia excelsa</i> Humb. & Bonpl.	–	1	–
18	Soybean <i>Glycine max</i> (L.) Merr., 1917	–	2	–
19	Pea <i>Pisum sativum</i> L., 1753	–	4	–
20	Perilla <i>Perilla frutescens</i> var. <i>hirtella</i> (Nakai) Makino	–	5	–
21	Almond <i>Prunus dulcis</i> (Mill.) D.A. Webb, 1967	–	8	4
22	Mandarin unshiu <i>Citrus unshiu</i> (Tanaka ex Swingle) Marcow., 1921	–	15	–
23	Tea <i>Camellia sinensis</i> (L.) Kuntze, 1887	–	–	1
24	Barley <i>Hordeum vulgare</i> L. (1753)	–	–	2
25	Hibiscus <i>Hibiscus syriacus</i> L. (1753)	–	–	2
26	Pineapple <i>Ananas comosus</i> (L.) Merr., 1917	–	–	3
27	Olive <i>Olea europaea</i> L., 1753	–	–	4
28	Sweet wormwood <i>Artemisia annua</i> L.	13	35	16
Total number of food plant species		15	12	12

1990) ↔ (iii) the TBP/promoter complex is stabilized by bending of the DNA double-helix axis at a right angle (Flatters, Lavery, 1998), as subsequently demonstrated experimentally *in vitro* (Delgadillo et al., 2009).

Statistical analysis. In this work, using standard software package Statistica (Statsoft™, USA), we averaged the Plant_SNP_TATA_Z-tester-generated (Rasskazov et al., 2022) estimates of K_D – for complexes of plant TBP with promoters of β -amylase, albumin, and globulin genes – for food and non-food plants separately. On the basis of these data, statistical significance of differences between food and non-food plants was evaluated by Fisher’s Z-test.

Results

Globulin

Table 3 presents the *in silico* estimates of K_D for complexes of plant TBP with 74 proximal promoters of globulin genes from 15 food plant species in comparison with 53 such promoters from 12 non-food plant species, as determined using Plant_SNP_TATA_Z-tester (Rasskazov et al., 2022). One can see in this table that in the case of food plants, the estimates of K_D for complexes of plant TBP with promoters of these genes varied from 1.67 ± 0.12 (mean \pm SEM) to 6.75 ± 5.23 nM, with an average of 2.97 ± 0.21 nM, whereas for non-food plants,

Table 2. Characteristics of 128 nucleotide sequences of proximal promoters from non-food plant genes homologous to the studied globulin (*Glo*), albumin (*Alb*), and β-amylase (*Bmy*) genes from common wheat *T. aestivum*

Non-food plant species		Number of promoters		
No.	Name	<i>Glo</i>	<i>Alb</i>	<i>Bmy</i>
1	Five-seeded plume-poppy <i>Macleaya cordata</i> (Willd.) R. Br.	1	–	–
2	Witchweed <i>Striga asiatica</i> (L.) Kuntze	1	1	1
3	Genlisea <i>Genlisea aurea</i> A.St. Hil. (1833)	1	–	2
4	Florida teosinte <i>Zea luxurians</i> (Durieu & Asch.) R.M. Bird, 1978	1	–	–
5	Noccidium <i>Microthlaspi erraticum</i> (Jord.) T. Ali & Thines, 2016	2	1	2
6	Gama grass <i>Tripsacum dactyloides</i> (L.) L., 1759	2	–	–
7	Balsas teosinte <i>Zea mays</i> subsp. <i>parviglumis</i> Iltis & Doebley, 1980	3	–	–
8	Thale cress <i>Arabidopsis thaliana</i> (L.) Heynh., 1842	4	5	3
9	Panic grass <i>Dichanthelium oligosanthos</i> (Schult.) Gould	6	–	6
10	Water lily <i>Nymphaea thermarum</i> Eb. Fisch., 1988	8	1	9
11	Rue-anemone <i>Thalictrum thalictroides</i> (L.) A.J. Eames & B. Boivin	9	15	6
12	Chile tomato <i>Solanum chilense</i> (Dunal) Reiche	15	13	6
13	Purple witchweed, <i>Striga hermonthica</i> (Delile) Benth.	–	1	–
14	Gerardia <i>Phtheirospermum japonicum</i> (Thunb.) Kanitz	–	–	1
15	Chinese rose <i>Rosa chinensis</i> Jacq., 1768	–	–	2
Total number of food plant species		12	7	10

these values varied from 1.25 ± 0.06 to 3.33 ± 0.23 nM, with an average of 2.15 ± 0.08 nM.

In Fig. 1, arithmetic mean estimates of K_D for complexes of plant TBP with globulin-coding gene promoters are compared between two groups (food and non-food plants) by Fisher's Z-test. The difference between the groups was significant, with $Z = 3.59$ and $p < 0.001$.

Albumin

Table 4 shows data obtained by Web service Plant_SNP_TATA_Z-tester (Rasskazov et al., 2022) regarding estimates of K_D for complexes of plant TBP with 84 albumin gene promoters from 12 food plant species and with 37 promoters from 7 non-food plant species. As readers can see in this table, in the case of food plants, the estimates of K_D of TBP-promoter complexes for these genes ranged between 1.65 ± 0.12 and 4.49 ± 1.39 nM (average: 3.10 ± 0.22 nM), whereas for non-food plants, they ranged from 1.65 ± 0.05 to 2.70 ± 0.22 nM (average: 2.18 ± 0.10 nM).

A comparison of the two groups (food and non-food plants) by Fisher's Z-test is displayed in Fig. 2. Here one can see a significant difference between food plants and non-food plants ($Z = 3.85$, $p < 0.001$).

β-Amylase

Table 5 lists estimated K_D values of complexes of plant TBP with 77 proximal promoters of β-amylase genes from 12 food

plant species and with 38 promoters from 10 non-food plant species, as calculated by Web service Plant_SNP_TATA_Z-tester (Rasskazov et al., 2022). For food plants, this table presents the range of K_D from 1.30 ± 0.09 to 8.77 ± 7.36 nM, with an arithmetic mean of 2.85 ± 0.21 nM, whereas for non-

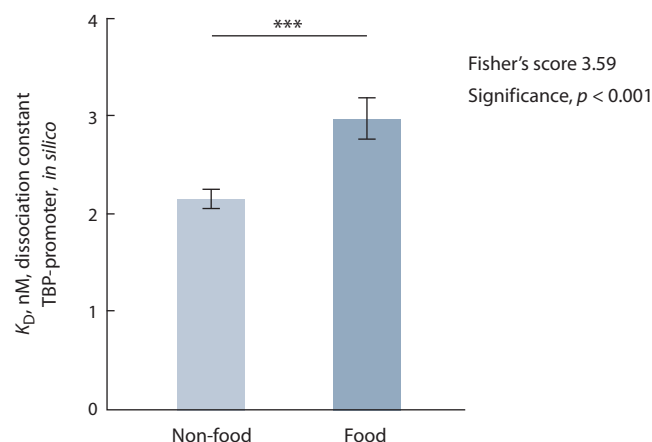


Fig. 1. The statistically significant difference between the studied food plants and non-food plants in the *in silico* estimates of K_D for complexes of plant TBP with 90 bp proximal promoters of their genes encoding globulins.

Here and in Fig. 2: *** statistical significance $p < 0.001$ according to Fisher's Z-test.

Table 3. Arithmetic mean estimates (M_0) of the equilibrium dissociation constant (K_D) of complexes between plant TBP and 90 bp proximal promoters of the plant globulin genes analyzed in this work

No.	Plant species	N	$K_D, M_0 \pm \Delta, \text{nM}$
Food plants			
1	Congolese coffee	1	2.17 ± 0.13
2	Buckwheat	1	2.04 ± 0.14
3	Maidenhair tree	1	1.67 ± 0.12
4	Yoshino cherry	1	1.76 ± 0.12
5	Maize	1	2.30 ± 0.16
6	Oat	2	2.57 ± 0.15
7	Waxberry	2	6.75 ± 5.23
8	Quinoa	2	2.47 ± 0.12
9	Rice	3	2.84 ± 0.38
10	Melon	4	2.66 ± 0.30
11	Cardoon	4	3.22 ± 0.37
12	Cork oak	4	4.84 ± 1.41
13	Wine grape	9	2.67 ± 0.35
14	Sweet wormwood	13	2.51 ± 0.25
15	Pepper	26	3.01 ± 0.37
Total		74	2.97 ± 0.21
Non-food plants			
1	Five-seeded plume-poppy	1	1.25 ± 0.06
2	Witchweed	1	3.33 ± 0.23
3	Genlisea	1	2.70 ± 0.19
4	Florida teosinte	1	1.96 ± 0.14
5	Noccidium	2	2.25 ± 0.83
6	Gama grass	2	2.19 ± 0.07
7	Balsas teosinte	3	2.12 ± 0.11
8	Thale cress	4	1.90 ± 0.13
9	Panic grass	6	2.27 ± 0.15
10	Water lily	8	2.60 ± 0.39
11	Rue-anemone	9	2.01 ± 0.16
12	Chile tomato	15	1.97 ± 0.11
Total		53	2.15 ± 0.08

Note. Here and in Tables 4 and 5: N – total number of the promoter studied; M_0 – arithmetic mean score; Δ – standard error of the mean (SEM).

food plants, the range of K_D was found to be 1.66 ± 0.32 to 6.75 ± 5.23 nM, with an average of 3.89 ± 0.32 nM. Fig. 3 presents a comparison between the analyzed food and non-food plants by Fisher’s Z-test, according to which these groups are statistically significantly different at $Z = 2.74$ and $p < 0.01$.

Table 4. Arithmetic mean estimates (M_0) of the equilibrium dissociation constant (K_D) for complexes between plant TBP and 90 bp proximal promoters of the plant albumin genes investigated in this work

No.	Plant species	N	$K_D, M_0 \pm \Delta, \text{nM}$
Food plants			
1	Sesame	1	2.28 ± 0.16
2	Kiwifruit nashi-kazura	1	1.77 ± 0.12
3	Brazil nut	1	3.04 ± 0.15
4	Waxberry	2	1.96 ± 0.14
5	Melon	2	2.04 ± 0.14
6	Soybean	2	1.65 ± 0.12
7	Pea	4	4.49 ± 1.39
8	Perilla	5	1.98 ± 0.40
9	Almond	8	3.74 ± 0.66
10	Pepper	8	3.00 ± 0.59
11	Mandarin unshiu	15	3.51 ± 0.61
12	Sweet wormwood	35	3.07 ± 0.35
Total		84	3.10 ± 0.22
Non-food plants			
1	Water lily	1	2.70 ± 0.22
2	Noccidium	1	2.00 ± 0.14
3	Purple witchweed	1	1.65 ± 0.12
4	Witchweed	1	2.19 ± 0.15
5	Thale cress	5	2.03 ± 0.21
6	Chile tomato	13	2.33 ± 0.20
7	Rue-anemone	15	2.11 ± 0.16
Total		37	2.18 ± 0.10

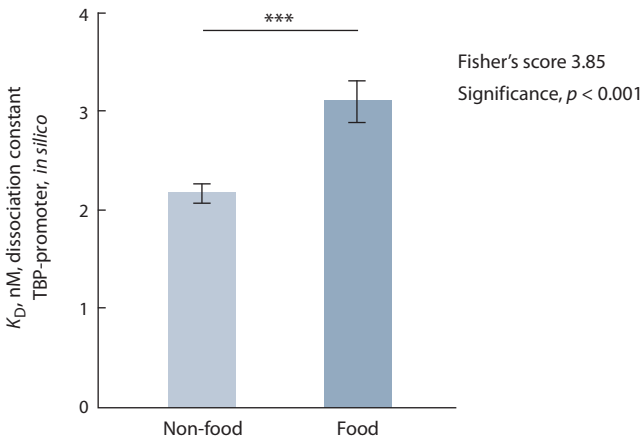


Fig. 2. The statistically significant difference between the studied food plants and non-food plants in the *in silico* estimates of K_D for the complexes of plant TBP with 90 bp proximal promoters of their genes encoding albumins.

Table 5. Arithmetic mean estimates (M_0) of the equilibrium dissociation constant (K_D) of complexes between plant TBP and 90 bp proximal promoters of the plant β-amylase genes examined in this work

No.	Plant species	N	$K_D, M_0 \pm \Delta, \text{nM}$
Food plants			
1	Tea	1	4.59 ± 0.28
2	Waxberry	1	2.21 ± 0.13
3	Barley	2	1.30 ± 0.09
4	Hibiscus	2	3.58 ± 1.86
5	Yoshino cherry	2	3.19 ± 1.68
6	Pineapple	3	8.77 ± 7.36
7	Almond	4	6.56 ± 1.63
8	Olive	4	5.24 ± 0.93
9	Melon	6	4.79 ± 0.96
10	Wine grape	9	3.97 ± 0.73
11	Sweet wormwood	16	4.29 ± 0.77
12	Pepper	27	2.59 ± 0.23
Total		77	3.89 ± 0.32
Non-food plants			
1	Witchweed	1	3.50 ± 0.25
2	Gerardia	1	3.43 ± 0.21
3	Genlisea	2	1.38 ± 0.82
4	Noccidium	2	1.93 ± 0.37
5	Chinese rose	2	1.79 ± 0.23
6	Thale cress	3	1.66 ± 0.32
7	Panic grass	6	3.26 ± 0.43
8	Rue-anemone	6	2.91 ± 0.43
9	Chile tomato	6	2.89 ± 0.47
10	Water lily	9	3.30 ± 0.62
Total		38	2.85 ± 0.21

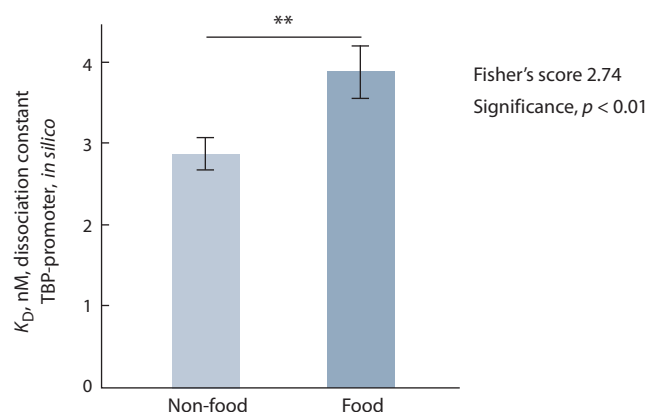


Fig. 3. The statistically significant difference between the studied food plants and non-food plants in the *in silico* estimates of K_D for the complexes of plant TBP with 90 bp proximal promoters of their genes encoding β-amylases.

** Statistical significance $p < 0.01$ according to Fisher's Z-test.

Discussion

It is well known that in the process of spontaneous domestication of ancestral forms of modern food plants, the selection was primarily based on their economically valuable traits, such as productivity, resistance to pathogens and to environmental stressors, and the ease of technological processing into final food products. Additionally, during the plant domestication, humans assessed the palatability of food products and their benefits for health.

It remains unclear whether in addition to the positive selection for the beneficial properties of agricultural plants, there was also simultaneous selection against their detrimental properties, which include allergenicity of the dishes prepared from these plants. To answer this question, we concentrated on the search for molecular genetic selection markers related to structural and functional organization of proximal promoters of plant genes.

Accordingly, plant genes were analyzed that are homologous to three *T. aestivum* genes encoding food allergens β-amylase, albumin, and globulin, earlier identified as targets for human IgE (Wang et al., 2021). Thus, 363 homologous genes were investigated belonging to 28 and 15 species of food plants and non-food plants, respectively. With the help of Web service Plant_SNP_TATA_Z-tester (Rasskazov et al., 2022), for each homologous gene, K_D of the complex of plant TBP with this gene's proximal promoter was computed.

Interest in the TBP protein and its binding site in the proximal promoter (canonical form: the TATA box) is due to the fact that they play a key role in the initiation of eukaryotic gene transcription. It has been experimentally established (Coleman, Pugh, 1995) that TBP slides along the DNA double helix owing to nonspecific affinity between them: $K_D \sim 10^{-5} \text{ M}$ (Hahn et al., 1989). TBP then stops at a site of TBP binding because of their mutual molecular recognition (Berg, von Hippel, 1987; Bucher, 1990) mediated by stronger (specific) affinity of TBP for this site: $K_D \sim 10^{-9} \text{ M}$ (Hahn et al., 1989). Next, under the action of TBP, the DNA double helix melts at the site of TBP binding, and kinking of the DNA axis at a right angle takes place, which stabilizes the TBP-promoter complex (Flatters, Lavery, 1998). The resultant TBP-promoter complex is considered an obligatory DNA anchor, which is required for the binding of RNA polymerase II (Muller et al., 2001; Martianov et al., 2002; Choukrallah et al., 2012; Rhee, Pugh, 2012) as a key step in the assembly of the transcription preinitiation complex (Auble, 2009) responsible for basal transcription (Fire et al., 1984). Due to the key importance of TATA boxes, mutations located in proximal promoters have a well-pronounced effect on the magnitude of gene expression (Savinkova et al., 2009).

The molecular mechanism underlying the binding of TBP to a promoter of various eukaryotic genes via the three successive steps was first proposed by P. Ponomarenko et al. (2008) and later confirmed experimentally (Delgadillo et al., 2009). Based on this mechanism, a bioinformatic model was devised previously for calculating a change in K_D (of a complex between TBP and a proximal promoter of a eukaryotic gene) for a polymorphism of the TBP-binding site(s) in the promoter as compared to the wild type (Ponomarenko et al., 2009). Results of computations based on this model have been

confirmed by independent *ex vivo* experiments on cell cultures transfected with the pGL4.10 plasmid (Promega, USA) carrying a wild-type or mutant promoter inserted before a luciferase reporter gene (Ponomarenko et al., 2017) as well as *in vitro* in real time (Arkova et al., 2017) by means of stopped-flow spectrometer SX.20 (Applied Photophysics, UK) under equilibrium conditions (Savinkova et al., 2013) and under non-equilibrium conditions (Drachkova et al., 2014) with the help of an electrophoretic mobility shift assay. As a result of such comprehensive verification of this bioinformatic model, on its basis, the Web service Plant SNP_TATA_Z-tester (Rasskazov et al., 2022) was created, which was employed in the current project for estimating K_D of complexes of plant TBP with proximal promoters of genes from food and non-food plants.

Our analysis revealed that in comparison with non-food plants, food plants are characterized by significantly weaker affinity of TBP for promoters of genes homologous to common-wheat β -amylase, albumin, and globulin (food allergens) ($p < 0.01$, as estimated by the above software and Fisher's Z-test). When interpreting the obtained results, let us take into account the experimentally proven fact that the level of expression of eukaryotic genes increases with enhancement of the affinity of TBP for the promoters of these genes (Mogno et al., 2010). This observation allows us to interpret the food plants' weaker TBP affinity – for promoters of genes homologous to genes of food allergens (common-wheat β -amylase, albumin, and globulin) in comparison with non-food plants – as evidence of selection by humans for low amounts of these allergenic proteins in food plants in the past, during the domestication of the plants.

Conclusion

In this work, DNA sequences of proximal promoters of genes homologous to genes of food allergens (Wang et al., 2021) were consistently analyzed *in silico* for the first time for food compared to non-food plants. As a result, weaker *in silico* affinity of TBP was observed for promoters of the investigated food plant genes as compared to genes of non-food plants. This finding is suggestive of artificial selection – in antiquity, for the purpose of reducing the expression of food plant genes encoding allergenic proteins – carried out by humans in the course of domestication of plants as food products.

References

- Arkova O., Kuznetsov N., Fedorova O., Savinkova L. A real-time study of the interaction of TBP with a TATA box-containing duplex identical to an ancestral or minor allele of human gene *LEP* or *TPI*. *J. Biomol. Struct. Dyn.* 2017;35(14):3070-3081. DOI 10.1080/07391102.2016.1241190.
- Auble D.T. The dynamic personality of TATA-binding protein. *Trends Biochem. Sci.* 2009;34(2):49-52. DOI 10.1016/j.tibs.2008.10.008.
- Benson D.A., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. *Nucleic Acids Res.* 2015;43(Database issue):D30-D35. DOI 10.1093/nar/gku1216.
- Berg O.G., von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 1987;193(4):723-750. DOI 10.1016/0022-2836(87)90354-8.
- 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.
- Cavazza A., Mattarozzi M., Franzoni A., Careri M. A spotlight on analytical prospects in food allergens: From emerging allergens and novel foods to bioplastics and plant-based sustainable food contact materials. *Food Chem.* 2022;388:132951. DOI 10.1016/j.foodchem.2022.132951.
- Choukrallah M.A., Kobi D., Martianov I., Pijnappel W.W., Mischerikow N., Ye T., Heck A.J., Timmers H.T., Davidson I. Interconversion between active and inactive TATA-binding protein transcription complexes in the mouse genome. *Nucleic Acids Res.* 2012;40(4):1446-1459. DOI 10.1093/nar/gkr802.
- Coleman R.A., Pugh B.F. Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J. Biol. Chem.* 1995;270(23):13850-13859. DOI 10.1074/jbc.270.23.13850.
- 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.
- Drachkova I., Savinkova L., Arshinova T., Ponomarenko M., Peltek S., Kolchanov N. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. *Hum. Mutat.* 2014;35(5):601-608. DOI 10.1002/humu.22535.
- Fire A., Samuels M., Sharp P.A. Interactions between RNA polymerase II, factors, and template leading to accurate transcription. *J. Biol. Chem.* 1984;259(4):2509-2516. DOI 10.1016/S0021-9258(17)43382-5.
- Flatters D., Lavery R. Sequence-dependent dynamics of TATA-Box binding sites. *Biophys. J.* 1998;75(1):372-381. DOI 10.1016/S0006-3495(98)77521-6.
- Hahn S., Buratowski S., Sharp P.A., Guarente L. Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. *Proc. Natl. Acad. Sci. USA.* 1989;86(15):5718-5722. DOI 10.1073/pnas.86.15.5718.
- Hong L., Pan M., Xie X., Liu K., Yang J., Wang S., Wang S. Aptamer-based fluorescent biosensor for the rapid and sensitive detection of allergens in food matrices. *Foods.* 2021;10(11):2598. DOI 10.3390/foods10112598.
- IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Biochem. J.* 1970;120(3):449-454. DOI 10.1042/bj1200449.
- Martianov I., Viville S., Davidson I. RNA polymerase II transcription in murine cells lacking the TATA binding protein. *Science.* 2002;298(5595):1036-1039. DOI 10.1126/science.1076327.
- 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.
- Muller F., Lakatos L., Dantonel J., Strahle U., Tora L. TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish. *Curr. Biol.* 2001;11(4):282-287. DOI 10.1016/S0960-9822(01)00076-8.
- Ponomarenko P., Chadaeva I., Rasskazov D.A., Sharypova E., Kashina E.V., Drachkova I., Zhechev D., Ponomarenko M.P., Savinkova L.K., Kolchanov N. Candidate SNP markers of familial and sporadic Alzheimer's diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front. Aging Neurosci.* 2017;9:231. DOI 10.3389/fnagi.2017.00231.
- Ponomarenko P.M., Ponomarenko M.P., Drachkova I.A., Lysova M.V., Arshinova T.V., Savinkova L.K., Kolchanov N.A. Prediction of the affinity of the TATA-binding protein to TATA boxes with single nucleotide polymorphisms. *Mol. Biol. (Moscow).* 2009; 43(3):472-479. DOI 10.1134/S0026893309030157.
- Ponomarenko P., Savinkova L., Drachkova I., Lysova M., Arshinova T., Ponomarenko M., Kolchanov N. 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.

- Prescott S.L., Logan A.C., Bristow J., Rozzi R., Moodie R., Redvers N., Haahntela T., Warber S., Poland B., Hancock T., Berman B. Exiting the anthropocene: achieving personal and planetary health in the 21st century. *Allergy*. 2022;77(12):3498-3512. DOI 10.1111/all.15419.
- Rasskazov D., Chadaeva I., Sharypova E., Zolotareva K., Khandaev B., Ponomarenko P., Podkolodny N., Tverdokhle N., Vishnevsky O., Bogomolov A., Podkolodnaya O., Savinkova L., Zemlyanskaya E., Golubyatnikov V., Kolchanov N., Ponomarenko M. Plant_SNP_TATA_Z-tester: a Web service that unequivocally estimates the impact of proximal promoter mutations on plant gene expression. *Int. J. Mol. Sci.* 2022;23(15):8684. DOI 10.3390/ijms23158684.
- Rhee H., Pugh B. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*. 2012;483(7389):295-301. DOI 10.1038/nature10799.
- 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.
- Savinkova L.K., Ponomarenko M.P., Ponomarenko P.M., Drachkova I.A., Lysova M.V., Arshinova T.V., Kolchanov N.A. TATA box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry (Moscow)*. 2009;74(2):117-129. DOI 10.1134/s0006297909020011.
- Wang Y., Weng J., Zhu C., Ai R., Zhou J., Wang C., Chen Q., Fu L. Allergenicity assessment and allergen profile analysis of different Chinese wheat cultivars. *World Allergy Organ. J.* 2021;14(7):100559. DOI 10.1016/j.waojou.2021.100559.

ORCID ID

O.V. Vishnevsky orcid.org/0000-0002-0347-2252
I.V. Chadaeva orcid.org/0000-0002-2724-5441
E.B. Sharypova orcid.org/0000-0002-1467-9312
P.M. Ponomarenko orcid.org/0000-0003-2715-9612
N.L. Podkolodny orcid.org/0000-0001-9132-7997

D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
O.A. Podkolodnaya orcid.org/0000-0003-3247-0114
L.K. Savinkova orcid.org/0000-0003-4543-4104
E.V. Zemlyanskaya orcid.org/0000-0002-4304-1190
M.P. Ponomarenko orcid.org/0000-0003-1663-318X

Acknowledgements. Russian Science Foundation grant No. 20-14-00140 supported this study. The authors are thankful to the multi-access Center "Bioinformatics" for the use of computational resources as supported by Russian government project FWNR-2022-0020 and the Russian Federal Science and Technology Program for the Development of Genetic Technologies.

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

Received September 16, 2022. Revised November 29, 2022. Accepted November 30, 2022.

FastContext: A tool for identification of adapters and other sequence patterns in next generation sequencing (NGS) data

E. Viesná^{1, 2}, V. Fishman^{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

 minja@bionet.nsc.ru

Abstract. The development of next generation sequencing (NGS) methods has created the need for detailed analysis and control of each protocol step. NGS library preparation protocols may include steps with incorporation of various service sequences, such as sequencing adapters, primers, sample-, cell-, and molecule-specific barcodes. Despite a fairly high level of current knowledge, during the protocol development process researchers often have to deal with various kinds of unexpected experiment outcomes, which result either from lack of information, lack of knowledge, or defects in reagent manufacturing. Detection and analysis of service sequences, their distribution and linkage may provide important information for protocol optimization. Here we introduce FastContext, a tool designed to analyze NGS read structure, based on sequence features found in reads, and their relative position in the read. The algorithm is able to create human readable read structures with user-specified patterns, to calculate counts and percentage of every read structure. Despite the simplicity of the algorithm, FastContext may be useful in read structure analysis and, as a result, can help better understand molecular processes that take place at different stages of NGS library preparation. The project is open-source software, distributed under GNU GPL v3, entirely written in the programming language Python, and based on well-maintained packages and commonly used data formats. Thus, it is cross-platform, may be patched or upgraded by the user if necessary. The FastContext package is available at the Python Package Index (<https://pypi.org/project/FastContext>), the source code is available at GitHub (<https://github.com/regnveig/FastContext>).

Key words: next generation sequencing; NGS; adapters; patterns search; read analysis.

For citation: Viesná E., Fishman V. FastContext: A tool for identification of adapters and other sequence patterns in next generation sequencing (NGS) data. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):806-809. DOI 10.18699/VJGB-22-97

FastContext: инструмент для контекстного анализа последовательностей в данных секвенирования нового поколения (NGS)

Э. Весна^{1, 2}, В.С. Фишман^{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

 minja@bionet.nsc.ru

Аннотация. Бурное развитие методов секвенирования нового поколения (next generation sequencing, NGS) породило потребность в детальном анализе и контроле качества на каждом этапе протокола приготовления геномных библиотек. Протоколы могут включать в себя этапы с внедрением различного рода служебных последовательностей, таких как адаптеры, праймеры, а также баркоды, специфичные для каждого образца, клетки или молекулы ДНК. Несмотря на достаточно высокий уровень современных знаний в молекулярной биологии, в процессе разработки протоколов NGS исследователи часто сталкиваются с неожиданными экспериментальными данными, которые могут быть результатом недостатка информации о молекулярных процессах, сопровождающих приготовление геномных библиотек, или, в отдельных случаях, дефектом производства реактивов. Обнаружение и анализ распределения служебных последовательностей в полученных молекулах ДНК могут быть важным источником информации, необходимой для оптимизации протокола приготовления геномных библиотек. В настоящей статье представлена утилита FastContext, с помощью которой возможен анализ структуры прочтений с точки зрения присутствия определенных последовательностей и их взаимного расположения в прочтении. Алгоритм принимает на вход необработанные данные секвенирования в формате FastQ, а затем генерирует удобные для интерпретации представления структуры прочтений на основе заданных пользователем паттернов, высчитывает количество подобных структур и их долю от общего числа прочтений. Несмотря на простоту алгоритма, FastContext может быть полезен при анализе структуры прочтений, он помогает лучше

понять молекулярные процессы, происходящие на разных стадиях приготовления геномных библиотек и, как следствие, открывает возможности для усовершенствования протокола. FastContext – это проект с открытым исходным кодом, распространяемый под свободной лицензией GNU General Public License v3, полностью написанный на языке программирования Python и основанный на широко используемых программных пакетах и форматах данных. Таким образом, он может быть легко использован под любой операционной системой, исправлен и дополнен при необходимости. FastContext доступен в виде пакета в Python Package Index (<https://pypi.org/project/FastContext>), исходный код хранится на GitHub (<https://github.com/regneig/FastContext>).

Ключевые слова: секвенирование нового поколения; NGS; адаптеры; поиск паттернов; анализ прочтений.

Introduction

Since the advent of next generation sequencing (NGS) methods 20 years ago, those methods have been actively evolving and are currently applied to various areas of biology. Due to the increasing capacity of sequencers, it is now possible to obtain billions of short molecule sequences in a single NGS run. In order to utilize such a high throughput of modern sequencers, there is a practice of sample pooling. This method requires incorporation of sample-specific service sequences (barcodes), which allow to distinguish individual samples in raw sequencing data.

Other types of service sequences could be incorporated into the target molecules, such as sequencing adapters and primers, biotin-labeled oligonucleotides for target molecules enrichment (Gridina et al., 2021), molecule- and cell-specific barcodes, which are designed to identify a molecule (Smirnov et al., 2020) and/or a cell of origin (Aldridge, Teichmann, 2020).

There are many strategies in molecular genetics that are used for service sequences incorporation: direct ligation of DNA or RNA molecules, template-switching activity of reverse transcriptases, and incorporation of synthetic DNA transposons. During the whole process of new NGS methods development it is crucial to control each protocol step. In light of that, detection and analysis of service sequences distribution may provide important information for protocol optimization.

Here we introduce the FastContext tool, which is designed to analyze and compute statistics on NGS read structures. FastContext allows to search for user-specified sequences in NGS reads, gather data on their linkage, frequency of occurrence, and present statistics in a user-friendly manner.

Materials and methods

The script is completely written in the programming language Python (version 3.8). It is packaged as a part of the Python Package Index (<https://pypi.org/project/FastContext>) and can be installed via pip. Therefore, it works out of the box on every operating system.

We used the following Python libraries:

1. bioPython, version 1.79 (Cock et al., 2009): FastQ files parsing and sequences manipulation;
2. python-Levenshtein¹, version 0.12.2: calculating sequences Levenshtein distance;
3. pandas, version 1.2.5 (The Pandas Development Team, 2020): tables creation;
4. tqdm, version 4.61.2 (Costa-Luis et al., 2022): visualization.

All libraries listed above, except python-Levenshtein, are widely used and well maintained.

FastContext supports multi-processing.

¹ Available at: <https://github.com/ztane/python-Levenshtein>.

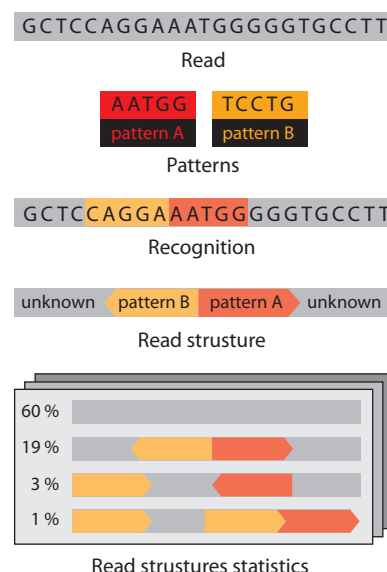


Fig. 1. FastContext algorithm scheme.

Two different example patterns colored as red and yellow.

Results

We developed an algorithm which parses raw sequencing dataset, searches each read or read pair for user specified patterns, and then generates a human-readable representation of the search results, which we call “read structure”. Algorithm scheme is represented in the Fig. 1.

Input

Input files are provided in FastQ² format. The user can provide one (in the single-end mode) or two (in the paired-end mode) FastQ files. Files may be uncompressed or compressed with gzip or bz2 algorithms.

Output

Output results are provided as an HTML page (further: “summary file”), containing run options and tables with read structures, their counts, and percentages (Fig. 2). The sequence strand (forward F, or reverse R) is displayed after a colon (e.g., {**oligb:F**}).

The user can manually set minimal rate value (rate floor) to be displayed. Also, the user can save the read structure for each read or read pair, with the read name, the sequence, and Phred qualities, as a gzip-compressed JavaScript Object Notation (JSON)³ object (further: “detailed statistics file”).

² Full specification of FastQ format is available at <http://maq.sourceforge.net/fastq.shtml>.

³ Full specification of JSON format could be found at JSON official website: <https://www.json.org>.

Count	Percentage	Read structure
5,197	48.80	{unknown}
3,297	30.96	{unknown}--{oligme:F}--{oligb:F}--{701:F}--{unknown}
114	1.07	{unknown}--{oligb:F}--{701:F}--{unknown}
71	0.66	{unknown}--{oligme:F}--{unknown}
69	0.64	{unknown}--{oligme:F}--{unknown}--{701:F}--{unknown}
60	0.56	{unknown}--{oligme:F}--{oligb:F}--{701:F}--{kmer:14bp}

Fig. 2. Example of statistics table.
Every fragment of read structure, except palindromic or unrecognized sequences, has a strand suffix. Short unrecognized sequences (K-mers) have a length suffix.

Patterns

Pattern names and sequences are provided as a plain JSON object, e. g.:
{**"foo"**: **"CTGTCTCTTATACAC"**, **"bar"**: **"CCGAAAACACG"**,
"baz": **"TCGTCGGG"**}.

It should be noted that pattern sequences are searched in the order they are provided by the user, forward strand (the sequence itself) first, reverse strand (a reverse complement of the sequence) after. Therefore, the patterns order matters in search and should be carefully considered before running the program. FastContext expects patterns to be sorted from long to short, which is the best option for overlapping or nested sequences, and otherwise gives you a warning.

K-mers

FastContext performs the search based on full match, and a pattern sequence with one single sequencing error will be skipped as an unrecognized sequence (alias {unknown}). This is especially important for long patterns, which are under-represented due to higher cumulative frequency of sequencing errors. In addition, oligonucleotide synthesis errors and some enzymatic steps of NGS library preparation, such as A-tailing, may produce molecules one base pair shorter or longer than expected. In order to simplify identification of such extended or truncated sequences, we have implemented the ability to mark short unrecognized sequences (K-mers) of certain length (e. g., {**kmer:14bp**}). If a K-mer identified in the read is one base longer or shorter than a pattern sequence, we can suppose this K-mer is the pattern sequence, and test the hypothesis in a more detailed analysis of reads.

Levenshtein distances analysis

Additional features implemeted to account for sequencing errors include analysis of Levenshtein distances between different pattern sequences (pattern analysis), and between pattern sequences and read sequence. Pattern analysis is shown in the summary file, data on every single read can be found in the detailed statistics file.

Analysis of distances between pattern sequences can prevent pattern match or nesting, when sequences are confused with each other because of a few sequencing errors. Also, FastContext warns the user about palindromes and sequences that can become palindromic because of sequencing errors. This kind of sequence may affect statistics of forward-reverse orientation.

Analysis of distances between pattern and read sequences can show similarity of an unrecognized sequence and a pattern

sequence, so the user could suggest the real read structure even if FastContext fails to do that. All these data may be found in a detailed statistics file, with Levenshtein read analysis enabled (disabled by default).

System requirements and performance

By design, FastContext stores FASTQ reads in random access memory (RAM), therefore, the only system limitation is the RAM size. Tests we have performed show that 8 Gb RAM is enough for processing 10,000 reads, which is a high enough sample size for practical application of the tool.

There are two stages that determine the time taken for completing a task. Reading data from a physical storage (HDD, SSD, etc.) depends on the storage characteristics. Read analysis is parallelized and depends on the core number. We estimated FastContext performance characteristics on the laboratory computing server with 16 cores and 50 Gb RAM. The dependence of processing speed on process count matches the expected values. 10,000 of paired-end reads are processing for 2 seconds with 4 cores used, saving JSON increases that time to 6 seconds. With Levenshtein statistics, the same data are processing for 11 seconds, and 80 seconds are required to save JSON.

Code access

FastContext source code is available at GitHub (<https://github.com/regnveig/FastContext>) and is distributed under GNU General Public License v3.

Discussion

Despite the simplicity of the algorithm, FastContext may be useful in read structure analysis. It has an appealing combination of cutadapt (Martin, 2011) and FastQC (Andrews, 2010) features.

Recently, A. Bravo et al. (2021) presented a tool named 2FAST2Q, which has features similar to FastContext, including extracting and counting feature occurrences in FastQ files. Unlike FastContext, 2FAST2Q can search for frequent unknown sequences (so called extract and count mode), can handle sequence mismatches, takes into account base Phred qualities, and therefore provides more accurate statistics on feature counts. The qualitative difference of FastContext is that the tool can collect statistics on relative position of features in the read and features linkage.

There remains the problem of sequencing errors. The possibility of errors is directly dependent on sequence length. FastContext performs the search based on full match, there-

fore, under equal conditions, pattern sequences of greater length have a lower chance to be found, which may impact resulting statistics.

Similarity based on Levenshtein distance is a crude approximation to probability of presence of a particular sequence. It fails to take account of *in vitro* processes during library preparation and sequencing. This problem may be solved in future versions. As for now, the user can find Phred quality scores for each read in a detailed statistics file, and estimate analysis quality manually.

Another possible feature that can be discussed is wildcards (symbols which denote more than one canonical nucleobase). This feature may be implemented in future versions.

Conclusion

From all of the above, we can conclude that FastContext is effective as a tool for NGS data analysis, and could be a very useful source of information in the development of new molecular biology methods.

References

- Aldridge S., Teichmann S. Single cell transcriptomics comes of age. *Nat. Commun.* 2020;11(1):4307. DOI 10.1038/s41467-020-18158-5.
- Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Bravo A., Typas A., Veening J. 2FASTQ: A general-purpose sequence search and counting program for FASTQ files [preprint]. *BioRxiv*. 2021. DOI 10.1101/2021.12.17.473121.
- Cock P., Antao T., Chang J., Chapman B., Cox C., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., de Hoon M. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. DOI 10.1093/bioinformatics/btp163.
- Costa-Luis C., Larroque S., Altendorf K., Mary H., Korobov M., Yorav-Raphael N., Ivanov I., Bargull M., Rodrigues N., Chen G., Newey C., Zugnoni M., Pagel M., Dektyarev M., Rothberg A., Lee A., Panteleit D., Dill F., Kemenade H., McCracken J., Nordlund M., Nechaev N., Desh O. tqdm: A fast, Extensible Progress Bar for Python and CLI. *Zenodo*. 2022. DOI 10.5281/zenodo.595120.
- Gridina M., Mozheiko E., Valeev E., Nazarenko L., Lopatkina M., Markova Z., Yablonskaya M., Voinova V., Shilova N., Lebedev I., Fishman V. A cookbook for DNase Hi-C. *Epigenetics Chromatin*. 2021; 14(1):15. DOI 10.1186/s13072-021-00389-5.
- 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.
- Smirnov A., Fishman V., Yunusova A., Korablev A., Serova I., Skryabin B., Rozhdestvensky T., Battulin N. DNA barcoding reveals that injected transgenes are predominantly processed by homologous recombination in mouse zygote. *Nucleic Acids Res.* 2020;48(2):719-735. DOI 10.1093/nar/gkz1085.
- The Pandas Development Team. pandas-dev/pandas: Pandas. *Zenodo*. 2020. DOI 10.5281/zenodo.3509134.

ORCID ID

E. Viesná orcid.org/0000-0003-3480-3963
V. Fishman orcid.org/0000-0002-5573-3100

Acknowledgements. This work was supported by Russian Science Foundation, grant No. 22-14-00247. High-throughput computations required for FastContext testing were performed using the Collective usage center of the Institute of Cytology and Genetics SB RAS, 121031800061-7 (Mechanisms of genetic control of development, physiological processes and behavior in animals).

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

Received July 18, 2022. Revised September 2, 2022. Accepted September 7, 2022.

Unique or not unique? Comparative genetic analysis of bacterial O-antigens from the Oxalobacteraceae family

S.D. Afonnikova^{1, 2}✉, A.S. Komissarov³, P.D. Kuchur³

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 ITMO University, SCAMT Institute, St. Petersburg, Russia

✉ svetaafonnikova@gmail.com

Abstract. Many plants and animals have symbiotic relationships with microorganisms, including bacteria. The interactions between bacteria and their hosts result in different outcomes for the host organism. The outcome can be neutral, harmful or have beneficial effects for participants. Remarkably, these relationships are not static, as they change throughout an organism's lifetime and on an evolutionary scale. One of the structures responsible for relationships in bacteria is O-antigen. Depending on the characteristics of its components, the bacteria can avoid the host's immune response or establish a mutualistic relationship with it. O-antigen is a key component in Gram-negative bacteria's outer membrane. This component facilitates interaction between the bacteria and host immune system or phages. The variability of the physical structure is caused by the genomic variability of genes encoding O-antigen synthesis components. The genes and pathways of O-polysaccharide (OPS) synthesis were intensively investigated mostly for Enterobacteriaceae species. Considering high genetic and molecular diversity of this structure even between strains, these findings may not have caught the entire variety possibly presented in non-model species. The current study presents a comparative analysis of genes associated with O-antigen synthesis in bacteria of the Oxalobacteraceae family. In contrast to existing studies based on PCR methods, we use a bioinformatics approach and compare O-antigens at the level of clusters rather than individual genes. We found that the O-antigen genes of these bacteria are represented by several clusters located at a distance from each other. The greatest similarity of the clusters is observed within individual bacterial genera, which is explained by the high variability of O-antigens. The study describes similarities of OPS genes inherent to the family as a whole and also considers individual unique cases of O-antigen genetic variability inherent to individual bacteria.

Key words: O-antigen gene clusters; lipopolysaccharide genes; comparative analysis; O-antigen; Oxalobacteraceae; *Massilia*; *Collimonas*; *Janthinobacterium*; saccharide gene cluster.

For citation: Afonnikova S.D., Komissarov A.S., Kuchur P.D. Unique or not unique? Comparative genetic analysis of bacterial O-antigens from the Oxalobacteraceae family. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):810-818. DOI 10.18699/VJGB-22-98

Сравнительный генетический анализ О-антигенов бактерий семейства Oxalobacteraceae: уникальность или тривиальность?

С.Д. Афонникова^{1, 2}✉, А.С. Комиссаров³, П.Д. Кучур³

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

³ Национальный исследовательский университет ИТМО, Институт SCAMT, Санкт-Петербург, Россия

✉ svetaafonnikova@gmail.com

Аннотация. Многие растения и животные способны устанавливать симбиотические взаимоотношения с микроорганизмами, в том числе с бактериями. Специфика этих взаимодействий может приводить к разным последствиям для организма-хозяина. Взаимоотношения могут быть нейтральными, негативными либо выгодными для одной или обеих сторон. Примечательно, что взаимоотношения бактерия-хозяин не являются статичными: они могут изменяться в течение жизни организмов и в ходе их эволюции. Одной из структур, определяющих направление изменчивости, является О-антиген. В зависимости от особенностей его компонентов бактерия может избегать иммунного ответа со стороны организма-хозяина, становясь патогеном, либо устанавливать с хозяином мутуалистические отношения. О-антиген – это ключевой компонент наружной мембраны грамотрицательных бактерий. Этот компонент обеспечивает взаимодействие между бактериями и иммунной системой хозяина или фагами. Вариативность структуры О-антигенов тесно связана с изменчивостью генов, кодирующих компоненты его синтеза. Гены и пути синтеза О-антигенов наиболее детально изучены у бактерий из семейства Enterobacteriaceae. С учетом высокого генетического и молекулярного разнообразия этой структуры даже между штаммами эти результаты могут не отражать все разнообразие О-антигенов, представленное у немодельных видов. В настоящей работе проведен сравнительный анализ генов, участвующих в

синтезе О-антигена, для бактерий из семейства Oxalobacteraceae. В отличие от существующих исследований, преимущественно основанных на методе ПЦР, в нашей работе использован биоинформатический подход, а сравнение проведено не на уровне одиночных генов, а на уровне кластеров. Мы обнаружили, что в случае Oxalobacteraceae генетическая организация О-антигена представлена несколькими кластерами, находящимися на значительном удалении друг от друга в геноме бактерий. Наибольшее сходство кластеров наблюдалось внутри отдельных родов бактерий, что объясняется высокой изменчивостью О-антигенов. В работе описано сходство генов О-антигенов, присущее семейству в целом, а также рассмотрены отдельные уникальные случаи изменчивости их генетической структуры у отдельных бактерий.

Ключевые слова: кластеры генов О-антигена; гены липополисахарида; сравнительный анализ; О-антиген; Oxalobacteraceae; *Massilia*; *Collimonas*; *Janthinobacterium*; кластеры генов сахаридов.

Introduction

The Oxalobacteraceae family belongs to the Burkholderiales order of Proteobacteria. According to the Integrated Taxonomic Information System (www.its.gov) this family includes 55 verified species of 12 genera. Members of the Oxalobacteraceae family are stained negatively by Gram and presented in a wide range of habitats (outlined in Supplementary Materials, Table S1)¹. Species were found in soils, including grassland, volcanic and heavy metal polluted soils, in water and glaciers (Baldani et al., 2014). Some of them are free-living, others may form various relationships with plants. Symbiotic species (*Massilia*, *Herbaspirillum*) are known to exhibit plant growth-promoting features, and can be beneficial in agriculture (Ofek et al., 2012; Peta et al., 2019; Grillo-Puertas et al., 2021). Occasionally, these relationships lead to plant diseases, for example, red stripe and mottle stripe diseases (Tuleski et al., 2020). The negative effect depends on the environment conditions. Examples of opportunistic features are described for *Janthinobacterium* and *Herbaspirillum* genera. Some species can be found in clinical samples and act as opportunistic pathogens for humans (Dhital et al., 2020).

Beneficial effects from Oxalobacteraceae bacteria are related to agriculture and medicine. Farming industry utilizes these bacteria to improve plant growth. Mutualistic bacteria facilitate nitrogen assimilation to increase crops productivity. In medicine, bacterial lipopolysaccharides (LPS) can be used for vaccine development. This modern medicine development is called glycoconjugate vaccines. The methodology is already verified on the members of Enterobacteraceae family (Bazhenova et al., 2021) and can be scaled to other bacteria. Beyond vaccines, information related to LPS lies in biosensor systems. Systems are able to identify bacteria in samples based on their LPS composition, in particular O-antigens (Sannigrahi et al., 2020).

O-antigen became a convenient feature for serotyping due to its variability. Diversity of the oligopolysaccharide units and the selection of the host immune system directed at them highly contribute to the variability of O-antigens. In addition to this selection, there is the bacteriophage effect on the bacterial cell (Xi et al., 2019). All these factors explain the emergence of different serotypes within the same bacterial species.

O-antigen is one part of bacterial LPS. Lipopolysaccharides are a specific structures (plural form) binding to the outer membrane of Gram-negative bacteria. It consists of three parts that are linked to each other in a particular order: phospholipid anchored to the membrane (lipid A or endotoxin), core region and O-antigen repeats. Lipid A is the hydrophobic domain an-

choring LPS in the membrane. In chemical structure, lipid A is a phospholipid based on glucosamine. It forms the monolayer of the outer membrane. Lipid A is responsible for the toxicity of Gram-negative bacteria. The second component of LPS is the core part. The first and the second LPS components are synthesized on the cytoplasmic side of the inner membrane of the bacterial cell, after which they are transported by ABC transporters into the periplasmic space (Valvano, 2015). The third component of LPS is O-antigen, which is synthesized separately from the previous parts. In a periplasmic space, all parts of LPS are combined together, then the fully synthesized LPS is transported to the outer leaflet of the cell membrane (Doerrler, 2006).

The composition of LPS and its parts varies between different species and between strains (Caroff, Karibian, 2003). In some strains O-antigen can be absent, thus referred to as “rough” LPS, others containing it are “smooth” (Erridge et al., 2002). The O-antigen consists of a series of repeating oligosaccharide units. The length and composition of the monomers vary quite widely among strains (Perepelov et al., 2009). Repeats can be homodimers or heterodimers. In addition, units can be linked linearly or can create a branched structure (Liu et al., 2020).

Sugar nucleotides are basic molecules that form an O-antigen backbone. The most common can be divided into several groups:

- dTDP-sugars (*rfb/rml* genes);
- CDP-sugars (*ddh* genes);
- GDP-sugars (*man* genes, *gmd*, *col*);
- UDP-glucoses (*ugd*, *gla*, *galE*);
- UDP-N-acetylglucosamines (*gne*, *gna*, *fnl* and *mna* genes).

Other nucleotide sugar genes include *nna* genes (N-acetylneuraminic acid synthesis), *hdd* genes and *gmh* (LD-mannoheptose and DD-mannoheptose) and *dmh* genes of 6-deoxy-D-mannoheptose synthesis pathway (Samuel, Reeves, 2003). The O-antigen chain is assembled via glycosyltransferases, which are responsible for combinations of sugar nucleotides.

The mechanisms of generating O-antigen and flipping are described in two variants: Wzy-dependent pathway and ABC-transporter pathway. The former is predominant among better-characterized O-antigens. A third variant is the synthase-dependent pathway. Unfortunately, it is poorly described and has been observed rarely, for instance, in *Salmonella* species (Kalynych et al., 2014).

The initiation of all O-antigen synthesis pathways is a transfer of a sugar monophosphate to the undecaprenyl phosphate (Und-P) molecule, resulting in sugar-pyrophosphate-undecaprenyl (sugar-Und-PP). Sugar-Und-PP is able to accept further glycosylation reactions (Kalynych et al., 2014).

¹ Supplementary Materials 1–6 are available in the online version of the paper: <https://doi.org/10.5281/zenodo.7410337>.

Uniquely to the Wzy-dependent pathway, Und-P-linked units are polymerized by Wzy (*wzy* gene) and subsequently flipped via Wzx (*wzx* gene). The chain length is controlled by Wzz protein (*wzz*). The completed structure is ligated to the outer core region via WaaL O-antigen ligase encoded by *waaL* (*rfaL*) gene (Han et al., 2012).

On the contrary, the ABC-transporter pathway needs only a single initiation reaction per O-antigen chain. Moreover, the entire polymerization process via glycosyltransferases is carried out in the cytoplasm. Then the completely generated O-antigen-Und-PP molecule is flipped to the periplasmic space by an ABC transporter, which is encoded by *wzt* and *wzm* genes. Similarly to the previously characterized pathway, the O-antigen ligase protein WaaL connects it to the core-lipid A (Samuel, Reeves, 2003).

In view of the above described, O-antigen becomes a highly variable structure. This feature makes the O-antigen attractive to a wide range of researchers. Nevertheless, there are rather few studies on comparative analysis of O-antigens and their genetic structure between bacteria at the family level. Most publications are devoted to single pathogenic or potentially pathogenic bacteria and avoid features of free-living or mutualistic species.

Detection and study of O-antigens have been made possible by the emergence of several methods involving both experimental and bioinformatics analysis of bacterial data. One of the traditional methods belonging to the first group is the bacterial glycotyping method based on the somatic antigen. In 2020, E.T. Sumrall et al. (2020) proposed a new method for quantitative separation of O-antigens. It is based on the use of a set of recombinant proteins that can interact with bacterial envelope receptors and domains. Bacterial O-antigens can also be detected by serological and agglutination test methods using sera specific to somatic antigens (Thakur et al., 2018). Another way to study O-antigen composition is the polymerase chain reaction method, which is widely used to compare O-antigens in several bacteria.

The emergence and subsequent decrease in the cost of sequencing opened new ways of O-antigen studying. *In silico* analysis methods have significantly reduced the time required for data processing, and many routine processes have been automated. Extensive databases have appeared that lead to the O-antigens analysis of several bacteria at once. In comparison to traditional methods of O-antigen detection, *in silico* methods are able to revise taxonomy misunderstandings, identify more genes related to O-antigen biosynthesis and evaluate their environment in a short time. Predicted features can be then verified by traditional laboratory methods. On the example of an Oxalobacteraceae member called *Janthinobacterium* sp. SLB01 (Belikov et al., 2021), the taxonomy was revised by this combined approach.

Here we present comparative analysis of O-antigens for 20 genomes from the Oxalobacteraceae family. According to the query in UniprotKB “(protein_name: O-antigen) AND (taxonomy_id:75682)” there are only 456 genes whose proteins are annotated as O-antigen biosynthesis genes for this family. Our bioinformatics approach based on homologues search eliminates difficulties in gene annotation. We also shift from describing single genes to comparing O-antigens at the level

of their candidate gene clusters to broad information about the gene content of Oxalobacteraceae O-antigens.

Materials and methods

Data. Initial data was derived from NCBI databases and included 20 genomes. The main criterion of assembly selection was a rather high quality, that is, no more than ten contigs. The reason for such a criterion was to decrease the possibility of gene clusters being disrupted by unresolved sequences. Overall, we selected two *Collimonas* species (*C. arenae* and *C. fungivorans*), one species of genera *Herminiimonas* (*H. arsenitoxidans*), *Oxalobacter* (*O. formigenes*), and *Undibacterium* (*U. parvum*), four *Janthinobacterium* (*J. agaricidamnorum*, *J. lividum*, *J. svalbardensis*, *J. tractae*), two *Oxalicibacterium* (*O. faegigallinarum* and *O. flavum*) species and nine *Massilia* (*M. albidiflava*, *M. armeniaca*, *M. flava*, *M. oculi*, *M. plicata*, *M. putida*, *M. timonae*, *M. umbonata*, *M. violaceinigra*). Their RefSeq assembly accessions are presented in Supplementary Materials, Table S2.

Quality control and annotation. All 20 assemblies were additionally analyzed using QUAST tool, version 5.0.2 (Gurevich et al., 2013). The acceptable threshold number of contigs and scaffolds was eleven, only genomes with a lower number were selected. To obtain the most precise annotation, we used two annotation tools, Prokka version 1.14.6 (Seemann, 2014) and eggNOG version 2.1.6 (Huerta-Cepas et al., 2019).

Putative O-antigen genes search. Searching for genes coding components for O-antigen synthesis and processing based on their names was unproductive because of the abundance of various synonymous tags. Therefore, we used an approach based on orthology. All O-antigen related genes for *Escherichia coli* strains described in the paper (Iguchi et al., 2015) were obtained with their amino acid sequences and used as reference. We also added genes from the KEGG database, a pathway of O-antigen synthesis for *E. coli* <https://www.genome.jp/pathway/ecoi00541>. We additionally analyzed the O-antigen ligase gene *rfaL* (*waaL*), because it was shown that O-antigen may be absent in some bacteria (Kime et al., 2016). As *waaL* is essential for final stages of O-antigen processing for the majority of bacteria, its absence may be associated with a lack of OPS on the cell wall (Wang et al., 2010). This data consisted of gene sets for each serogroup and approximately 420 unique genes in total (Supplementary Materials, Table S3).

In order to find unique genes among this data, sequences were clustered using UCLUST (Edgar, 2010) algorithm with the usearch32 tool, with threshold identity > 0.4. The reason for the rather low threshold was the excessive amount of clusters at higher numbers, mainly because of high gene variation. For the next step, we chose centroids of each cluster as representative sequences.

To reveal genes that correspond to the processed O-antigen genes of *E. coli*, we used the tool Orthofinder (Emms, Kelly, 2019) (version 2.5.4), which is able to find orthogroups and orthologs. Centroids data was taken as a reference. We assign functions of *E. coli* reference genes to all the Oxalobacteraceae sequences that fall into the same orthologous group.

The gene cluster is defined as a set of genes involved in a common metabolic pathway located within the genomic region

of 27,000 bp in length (Cimerancic et al., 2014). However, another important parameter for our definition is genes on borders. Thus, for an array of three genes, if genes on the borders of the set coincide, we also define this set as a cluster. A more detailed investigation of the obtained gene clusters with respect to their structure, function and sequence similarity was conducted using eggNOG and BLAST (v.2.5.0+) tools.

Verification of the identified candidate genes was performed via functional Pfam domains search (Supplementary Materials, Table S4). Lists of domains were obtained manually, from (Iguchi et al., 2015; Pereira et al., 2015). The HMMER software hmmer.org version 3.3.2 allowed the detection of those domains in FASTA amino acid sequences of all genomes. Some genes were checked manually using online Pfam sequence search <https://pfam.xfam.org> (Mistry et al., 2021). Characterization of genes shown to be uninvolved in O-antigen biosynthesis processes was performed using KEGG databases (Kanehisa, 2000).

Phylogenetic tree reconstruction. Phylogenetic tree was constructed to explore evolutionary relationships between the chosen Oxalobacteraceae taxa. Several species of the Burkholderiaceae family were selected to create an outgroup (*Burkholderia sordidicola*, *B. unamae*, *B. symbiotica*, *Ralstonia pickettii*, *Cupriavidus necator*). 16S rRNA sequences for 13 Oxalobacteraceae species were derived from published papers (Lim et al., 2003; Caballero-Mellado et al., 2004; Zhang et al., 2006; Sheu et al., 2012; Baldani et al., 2014; Koh et al., 2017; Daniel et al., 2021; Jung et al., 2021). Barnap version 0.9 (RRID:SCR_015995) was used for seven other genomes (*C. arenae*, *C. fungivorans*, *J. agaricidamnosum*, *J. lividum*, *J. svalbardensis*, *M. timonae*, *O. flavum*) to derive 16S rRNA sequences (Supplementary Materials, Table S5).

16S rRNA sequences were aligned using R-coffee, the T-coffee web-server RNA sequences alignment tool (Notre-dame et al., 2000). This tool takes into consideration the RNA secondary structure. Default multiple alignment options were chosen. The resulting alignment was used for constructing a phylogenetic tree using IQ-TREE web server (Nguyen et al., 2015). DNA was selected for sequence type, other options remained default. The best-fit model was TN+I+G4, the tree constructed with the Maximum likelihood method. Consensus tree was constructed from 1000 bootstrap trees and branch lengths were optimized by Maximum likelihood on original alignment. The results were visualized using Archaeopteryx 0.9928 (Han, Zmasek, 2009).

Gene clusters visualization. To visualize the found clusters we developed a Python script based on the DnaFeaturesViewer library (<https://edinburgh-genome-foundry.github.io/DnaFeaturesViewer/index.html#more-biology-software>). The code is available on this page https://github.com/svetaafonnikova/O-antigen-project/blob/main/draw_cluster.py. All steps of the data analysis algorithm are schematically depicted in Fig. 1.

Results

Assembly quality characterization

Out of all 20 assemblies, 15 were at the level of complete genomes. *M. timonae* assembly consisted of a single contig with N50 equal to the length of this contig. Two assemblies contained plasmid sequences (*M. putida* and *M. violacei-*

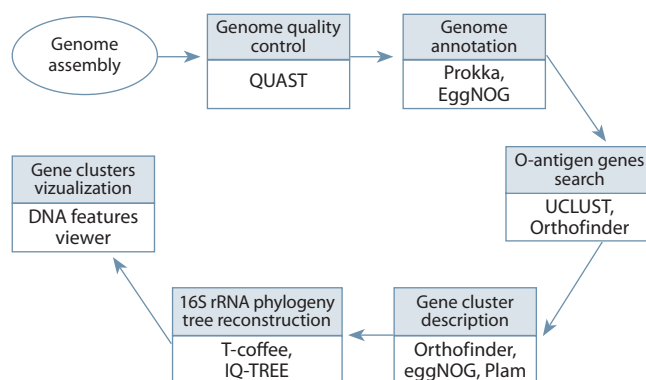


Fig. 1. Schematic representation of the data analysis algorithm used in the current study.

nigra). Another pair, *O. faecigallinarum* and *O. flavum*, contained ten and nine contigs, respectively.

Using IGV (v. 2.11.1) (Robinson et al., 2011) we confirmed that the identified O-antigen gene clusters were not located on plasmid fragments in case of plasmid containing genome assemblies. Secondly, O-antigen gene clusters were not situated on the borders of contigs, thus any breaks inside clusters were excluded.

Description of gene clusters

In general, almost all of the analyzed species contained more than two O-antigen gene clusters. These clusters are scattered around the genome and include not only O-polysaccharide genes, but genes of other functions. The visualization for all 20 species can be found in Supplementary Materials, Fig. S1. In the text below, we will describe these clusters for each genus used in the study.

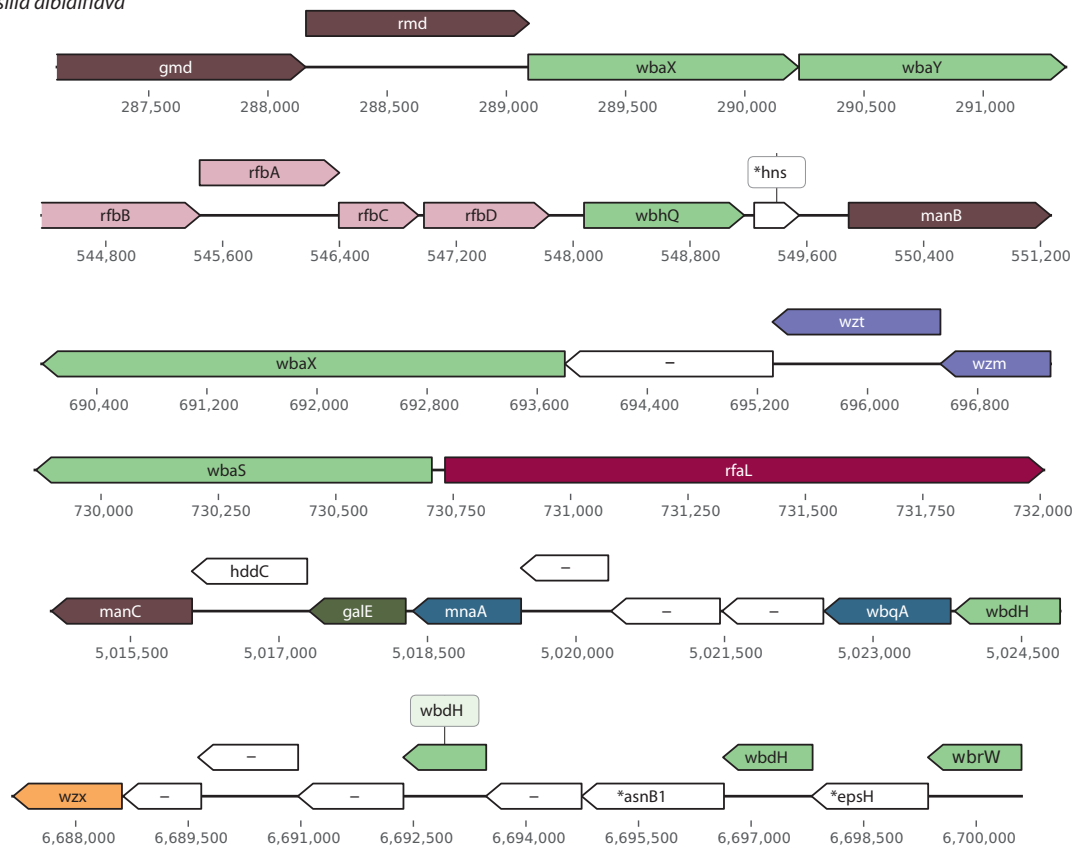
Collimonas. In both *C. arenae* and *C. fungivorans* we detected O-antigen ligase gene *rfaL* (or *waaL*) immediately adjacent to *galE* gene involved in nucleotide sugar synthesis. In addition, both genomes contain *wzm* and *wzt* genes. Furthermore, they share the same cluster with *manB* and *wfaK* on borders. All genes and their order coincide except one glycosyltransferase gene *wbaS*, absent in *C. fungivorans*.

Regarding other differences, the former species consists of three clusters, the latter consists of four. One of *C. fungivorans* clusters contains O-antigen unit synthesis (*rmd*, *gmd*, *manC*), processing genes (*wzm*, *wzt*) and a triplet of glycosyltransferase gene *wbaX*. Remarkably, in *C. arenae* these processing and unit synthesis genes are included in a single cluster with *rfaABCD* and *manB* genes on the borders.

Hermiimonas. According to our analysis, *H. arsenitoxidans* genome possesses three O-antigen gene clusters, with *rfaL* gene located outside all of them without any OPS genes beside. Regarding genes involved in processing, only *wzx* was observed. There are duplication instances for L-Rhamnose biosynthesis gene *rfaB*, sugar transferase genes *wbaT* and *wbaS*. One cluster contains a rather small number of genes we are interested in compared to not O-antigen ones. These unnecessary for OPS production genes partake in phosphate metabolism.

Janthinobacterium. *J. lividum* carries two clusters and *J. agaricidamnosum* comprises three gene clusters involved

Massilia albidiflava



Oxalobacter formigenes

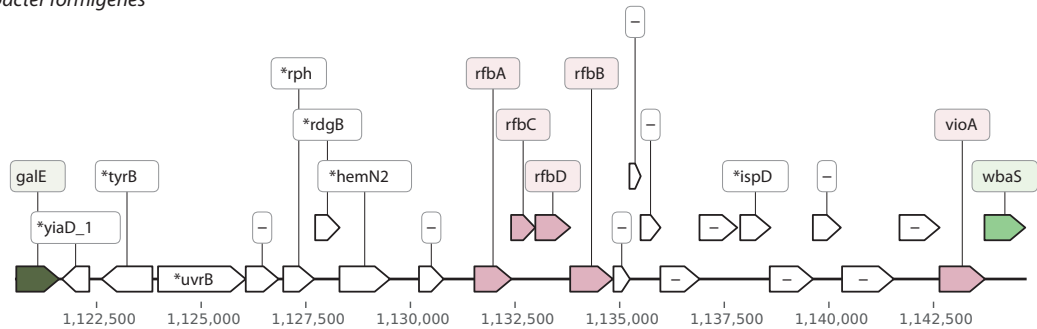


Fig. 2. The O-antigen gene clusters from *M. albidiflava* and *O. formigenes*. Unannotated genes are designated as “-”. Colors code for biosynthetic pathways. Orange genes are involved in Wzx/Wzy pathway, violet are involved in Wzm/Wzt pathway, rose genes partake in dTDP-sugar pathway, dark green, in UDP-sugar pathway, brown, in GDP-sugar pathway, *rfaL* gene is coded in red. UDP-N-Acetylglucosamine genes are blue, and transferase genes are light green. Genes involved in other pathways are white. The complete graphical visualization of OPS gene clusters for other analyzed species can be found in Supplementary Materials, Fig. S1.

in O-antigen synthesis, whereas *J. svalbardensis* and *J. tructae* include four. The latter two share identical clusters with UDP-N-acetylglucosamine pathway, *wbqA* and *wbqB* on the one end and glycosyltransferase gene *wbdH* on the other. All but *J. agaricidamnorum* have duplications of *rfbABCD* genes. All four genes are duplicated in *J. lividum* and *J. svalbardensis*, *J. tructae* possesses three copies of *rfbA* and *rfbB*. Furthermore, the *J. tructae* cluster with *rfbBA* and *fnlA* borders is almost similar to a part of another larger O-antigen gene cluster. In *J. lividum* and *J. svalbardensis* we found a common OPS related gene cluster flanked by *wbqB* and *wbhQ*. This gene set includes dTDP-glucose pathway genes *rfbABCD* and *vioA*. Still, the latter species has glycosyltransferase *wbaS*

next to *wbqB*, which *J. lividum* lacks in this position. To add, *wzx* gene was located after *vioA* in *J. lividum*, however, we didn't observe any significant domains for *J. svalbardensis* in that position. The O-antigen ligase was observed in all genus members. It lies far from any depicted cluster. Considering genes not included in our initial gene list, there are genes involved in LPS core synthesis (*waaD*), polysaccharide transport gene (*wza*), genes characteristic to O-antigen production in other bacteria species (*rfbG*, *rfbF*). **Massilia.** According to our analysis, *Massilia* is the genus with the highest number of O-antigen gene clusters. *M. oculi* has six clusters, *M. flava*, *M. umbonata* and *M. violaceinigra* possess only four and others contain five clusters (Fig. 2).

We observed some patterns in gene clusters between species. All *Massilia* species carry the *gmd_rmd_wbaX_wbaY* (in exact order) cluster. Only in *M. oculi* and *M. timonae* *rfbBDAC* genes (order in cluster) are presented as an independent cluster. In other genomes, these genes are surrounded by various O-antigen related genes. The same cluster with *rfbABCD* (order in cluster) genes and *manB* occurs in *M. violaceinigra*, *M. plicata*, *M. flava*, *M. armeniaca* and *M. albidiflava*. A single gene unrelated to O-antigen production is DNA-binding protein gene *hns*.

To add more similarity between *M. oculi* and *M. timonae*, they possess identical clusters consisting of *wbrW*, *wbdH*, *wbqB*, *ugd* on the one end and *wzx* on the other end. Genes located among them partake in infection initiation (*espH*), amino acid biosynthesis (*asnB1*), acyl-CoA and fatty acids biosynthesis (*fadD*).

In all assemblies, we observed *wzm* and *wzt* genes. Most of the species contain these genes in the order *wzm*, *wzt*, unannotated gene and *wbaX*. The group with such a set includes *M. albidiflava*, *M. armeniaca*, *M. oculi*, *M. putida*, *M. plicata*, and *M. timonae*. Another gene context is larger, the cluster is flanked by *wzm/wzt* and *vioA*. Between them are two glycosyltransferase genes *wbaX* with different lengths, unannotated genes and *gtrB*. The latter is a viral gene, and it can actually modify O-antigen structure. However, it was not described for the *E. coli* OPS gene cluster. Finally, in *M. violaceinigra* we found a unique set (not O-antigen biosynthesis gene cluster by our definition) of OPS processing genes and *wbaX*. There are three unannotated genes and two *wzt*. For the one beside *wzm* we didn't verify a specific domain, it was indicated as a gene not involved in O-antigen synthesis. The domain structure for *wzt* laying further from *wzt* was proved.

One of *M. armeniaca* clusters contains a full cluster described for *M. plicata*. It starts with *mnaA*, proceeds with *wbrW* and three copies of *wbdH*. In the former species, *wzx* with unannotated genes is added after the third *wbdH*. *M. umbonata* shares the most part of this cluster with *M. armeniaca*, except it lacks *mnaA* at the beginning. All genomes except *M. plicata* include the *galE*, *hddC* and *manC* part in the exact order in one cluster per genome.

Considering O-antigen ligase gene *rfaL*, in *M. albidiflava*, *M. oculi*, *M. plicata*, *M. timonae* and *M. violaceinigra* this gene is located next to *wbaS*. The rest of the species contain *rfaL* outside O-antigen clusters.

It can be noticed that some genes, for instance, *wbaS*, *rfaA* and *rfaB*, *mnaA*, *wbdH*, are presented in two or more copies in genomes.

Oxalicibacterium. Three clusters were identified for each species of the *Oxalicibacterium* genus. They share a cluster flanked by *wfaK* and *manC*. Their content slightly diverges from each other. *O. flavum* has more genes, including an additional O-antigen related gene *ugd*. The OPS ligase gene *rfaL* was identified in both assemblies, however, they are located in different contexts.

Oppositely to *O. flavum*, *O. faecigallinarum* carries UDP-N-Acetylglucosamine pathway genes (*fnlA*, *fnlB*, *mnaA*, *gne*, *wbqB*). On top of it, in the *O. faecigallinarum* we could locate duplications of the *rfbABCD* part, lying in discrete clusters and ordered in a different manner. However, *rfbD* gene in the

bigger cluster is rather dubious, the smaller length compared to other *rfbD* instances adds more uncertainty. We did not find this gene using Orthofinder analysis, although there is a Pfam domain corresponding to typical *rfbD* (RmlD_sub_bind) and it was annotated as *rfbD* by EggNOG.

We could detect *wzt/wzm* genes only in *O. flavum* assembly. The second species probably either does not carry these genes or they can be located outside clusters in unread spaces between contigs.

Oxalobacter. For *O. formigenes* we identified a single OPS cluster carrying dTDP-sugar pathway genes *rfbABCD* and *vioA* and UDP-glucose synthesis gene *galE*. The rest of the genes in the cluster are involved in nucleotide metabolism and cofactor synthesis. Also, any O-antigen processing genes were undiscovered. We couldn't detect *rfaL* gene in the given assembly. Moreover, even NCBI databases don't have any information considering this gene or protein in the *Oxalobacter* genus.

Undibacterium. For *U. parvum* two clusters were identified, *wzt* and *wzm* genes, were located outside them. Interestingly, *wzt* gene is smaller in comparison to this gene's length in other Oxalobacteraceae species. Typical of them, *wzt* is longer than *wzm* by approximately 400 bp. In contrast, *U. parvum*'s *wzt* is almost the same size as *wzm*. Using Pfam service, the gene's domain (ABC_tran) was verified.

Both clusters possess transferase and nucleotide sugar genes. Most spaces between OPS synthesis genes are unannotated genes, except *dyp* (peroxidase) and *ansA* (asparaginase) genes. The cluster carrying *rfbABCD* genes has a copy of *manC* gene and two *wbaX* genes, which have different sizes.

Phylogenetic tree

The phylogenetic dendrogram based on 16S rRNA showed that the chosen species clustered together considering their genera (Fig. 3). There had been no study including all species and their exact strains used in the current work. Therefore, we could compare only some clades of the tree. Similar to other studies, the first species to branch off is *Oxalobacter* species. Contrary to literature reports, our tree has a distinct *Oxalicibacterium* group and *Collimonas* with the rest of the species of the Oxalobacteraceae family (Baldani et al., 2014). However, the bootstrap support is rather small at this node. The *Janthinobacterium* group formation coincided with other papers (Jung et al., 2021). Some *Massilia* species clustered according to literature (Feng et al., 2016; Ren et al., 2018). Also, we obtained an unresolved node between *M. armeniaca* and *M. plicata*. The gene lengths used in the analysis varied between 1400 and 1500 bp for most cases (see Supplementary Materials, Table S5).

Discussion

In this work, we determined candidate genes involved in O-antigen biosynthesis in bacteria from the Oxalobacteraceae family. In comparison to well-studied *E. coli* O-antigen genes, they are presented in the form of several clusters. A similar situation has already been described for non-model bacteria (Hug et al., 2010). These clusters are dispersed across the genome. Clusters include O-antigen genes together with additional genes, which are necessary for LPS biosynthesis (for example, for core part synthesis and LPS parts binding) or

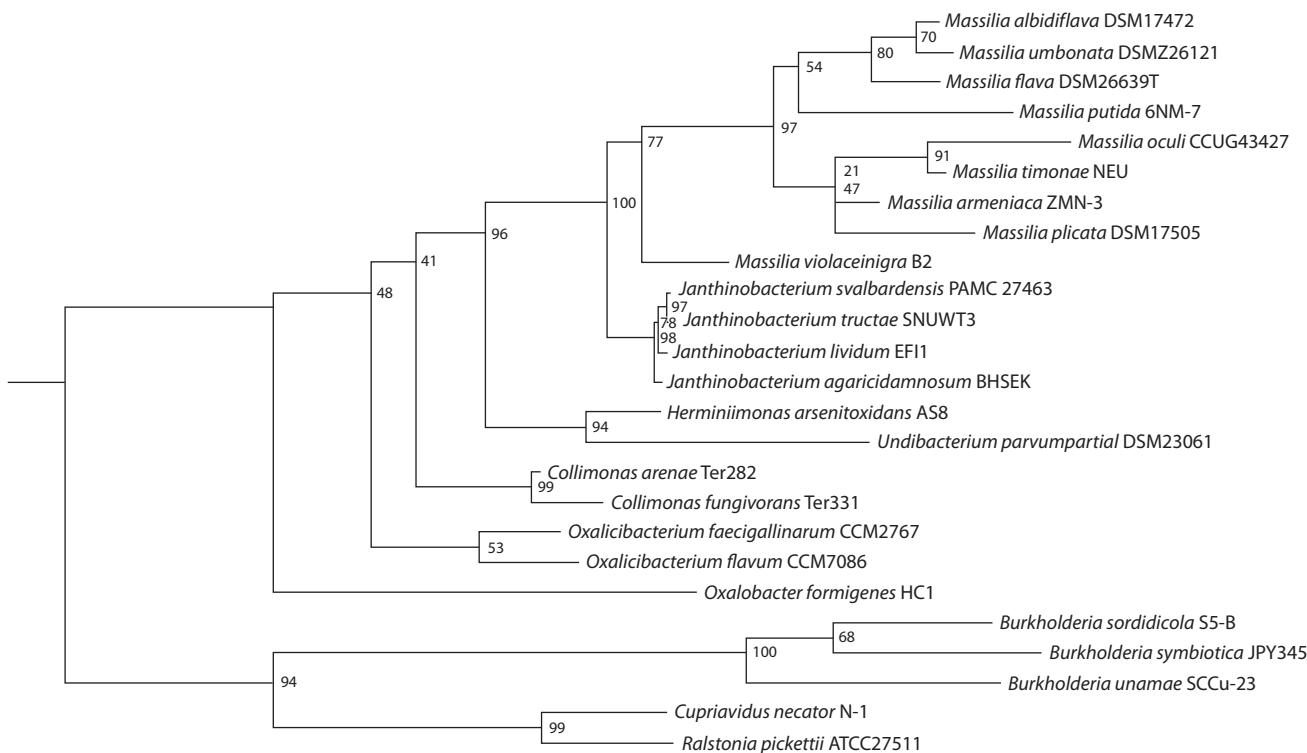


Fig. 3. Phylogenetic reconstruction of Oxalobacteraceae family members selected for the study based on 16S rRNA and created using Maximum likelihood method.

The consensus tree was obtained from 1000 bootstrap trees. The sequence data is described in Supplementary Materials, Table S3.

partake in other processes. The *E. coli* O-antigen gene cluster was studied by traditional laboratory methods, in particular, by PCR (DebRoy et al., 2011; Iguchi et al., 2015). These methods aim to detect specific genes, whereas *in silico* methods take into account the gene environment. In other words, they allow structures to be studied at the cluster level. Thus, our approach helps to expand understanding of the O-antigen genetic composition in bacterial genomes.

During OPS genetic structure comparison, we identified common features presented in all species inside the Oxalobacteraceae family. In particular, the group of *rfaABCD* genes was detected in each bacterium. The order of these genes varies, however, they are always placed together in one cluster. No one gene has deletions, nonsense mutations and other sequence abnormalities. According to the results, the studied bacteria should have a correct dTDP-rhamnose synthesis.

More similarities were found within each genus. These similarities relate mainly to individual genes or pairs of genes. A possible explanation lies in the high level of variability of O-antigens and the rate of bacterial mutations. O-antigens undergo changes so frequently that most of the similarities occur at the species or lower levels rather than at the genus or family level (Liu et al., 2008).

In 13 bacteria species, *wzm* and *wzt* genes were detected. We consider the Wzm-Wzt transporters pathway as the main path of O-antigen biosynthesis in this case (Wang et al., 2010). Wzx-Wzy pathway was not confirmed due to the absence of *wzy* genes.

Another interesting finding concerns gene duplication. The most repetitive genes were identified in *Massilia* species (see

Fig. 2). Its O-antigens clusters may contain up to three copies of the same gene. We suggest two possible explanations. The first one is related to the biological features of LPS. The same gene can provide the synthesis of several parts of LPS. The appearance of additional gene copies can increase the amount of protein in the cell or maintain its level in case one of the gene copies is broken. The second explanation is linked with an algorithm of O-antigen genes search. In our approach, genes are detected according to the principle of homology, so similar genes can be assigned the same name.

Symbiotic bacteria *Oxalobacter formigenes* lacks O-antigen ligase gene (*walL*) in O-antigen clusters, which may indicate the absence of O-antigen. The lack of the mentioned structure was discussed by J.K. Kim et al. (2016) for *Burkholderia* bacteria species. With our results, we confirm the possibility of loss of O-antigen genes in symbiotic bacterial species.

Conclusion

Overall, the findings of this study indicate differences of non-model bacteria from the model one by the example of the Oxalobacteraceae family. We suggest that the characterized OPS gene cluster composition is atypical. So far, most papers, which explored these genes for other bacteria, described only a single gene cluster. The O-antigen genetics of non-model bacteria is highly diverse, which is proved by the bioinformatic approach. The search for homologous sequences allows us to expand and deepen our understanding of gene clusters involved in O-antigen biosynthesis. Further investigation of the Oxalobacteraceae O-antigen genetic composition can be confirmed by laboratory methods.

References

- Baldani J.I., Rouws L., Cruz L.M., Olivares F.L., Schmid M., Hartmann A. The family *Oxalobacteraceae*. In: Rosenberg E., DeLong E.F., Lory S., Stackebrandt E., Thompson F. (Eds.). The Prokaryotes. Berlin; Heidelberg: Springer, 2014;919-974. DOI 10.1007/978-3-642-30197-1_291.
- Bazhenova A., Gao F., Bolgiano B., Harding S.E. Glycoconjugate vaccines against *Salmonella enterica* serovars and *Shigella* species: existing and emerging methods for their analysis. *Biophys. Rev.* 2021;13(2):221-246. DOI 10.1007/s12551-021-00791-z.
- Belikov S.I., Petrushin I.S., Chernogor L.I. Genome analysis of the *Janthinobacterium* sp. strain SLB01 from the diseased sponge of the *Lubomirskia baicalensis*. *Curr. Issues Mol. Biol.* 2021;43(3):2220-2237. DOI 10.3390/cimb43030156.
- Caballero-Mellado J., Martínez-Aguilar L., Paredes-Valdez G., Estrada-de los Santos P. *Burkholderia unamae* sp. nov., an N₂-fixing rhizospheric and endophytic species. *Int. J. Syst. Evol. Microbiol.* 2004; 54(4):1165-1172. DOI 10.1099/ijs.0.02951-0.
- Caroff M., Karibian D. Structure of bacterial lipopolysaccharides. *Carbohydr. Res.* 2003;338(23):2431-2447. DOI 10.1016/j.carres.2003.07.010.
- Cimercancic P., Medema M.H., Claesen J., Kurita K., Wieland Brown L.C., Mavrommatis K., Pati A., Godfrey P.A., Koehrsen M., Clardy J., Birren B.W., Takano E., Sali A., Linington R.G., Fischbach M.A. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell.* 2014;158(2):412-421. DOI 10.1016/j.cell.2014.06.034.
- Daniel S.L., Moradi L., Paiste H., Wood K.D., Assimios D.G., Holmes R.P., Nazzari L., Hatch M., Knight J. Forty years of *Oxalobacter formigenes*, a gutsy oxalate-degrading specialist. *Appl. Environ. Microbiol.* 2021;87(18):e0054421. DOI 10.1128/AEM.00544-21.
- DeRoy C., Roberts E., Frattamico P.M. Detection of O antigens in *Escherichia coli*. *Anim. Heal. Res. Rev.* 2011;12(2):169-185. DOI 10.1017/S1466252311000193.
- Dhital R., Paudel A., Bohra N., Shin A.K. *Herbaspirillum* infection in humans: a case report and review of literature. *Case Rep. Infect. Dis.* 2020;2020:9545243. DOI 10.1155/2020/9545243.
- Doerfler W.T. Lipid trafficking to the outer membrane of Gram-negative bacteria. *Mol. Microbiol.* 2006;60(3):542-552. DOI 10.1111/j.1365-2958.2006.05130.x.
- Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010;26(19):2460-2461. DOI 10.1093/bioinformatics/btq461.
- Emms D.M., Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. DOI 10.1186/s13059-019-1832-y.
- Erridge C., Bennett-Guerrero E., Poxton I.R. Structure and function of lipopolysaccharides. *Microbes Infect.* 2002;4(8):837-851. DOI 10.1016/S1286-4579(02)01604-0.
- Feng G.-D., Yang S.Z., Li H.P., Zhu H.H. *Massilia putida* sp. nov., a dimethyl disulfide-producing bacterium isolated from wolfram mine tailing. *Int. J. Syst. Evol. Microbiol.* 2016;66(1):50-55. DOI 10.1099/IJSEM.0.000670.
- Grillo-Puertas M., Villegas J.M., Pankiewicz V.C.S., Tadra-Sfeir M.Z., Teles Mota F.J., Hebert E.M., Brusamarello-Santos L., Pedraza R.O., Pedrosa F.O., Rapisarda V.A., Souza E.M. Transcriptional responses of *Herbaspirillum seropedicae* to environmental phosphate concentration. *Front. Microbiol.* 2021;12:666277. DOI 10.3389/FMICB.2021.666277.
- Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072-1075. DOI 10.1093/bioinformatics/btt086.
- Han M.V., Zmasek C.M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics.* 2009;10:356. DOI 10.1186/1471-2105-10-356.
- Han W., Wu B., Li L., Zhao G., Woodward R., Pettit N., Cai L., Thon V., Wang P.G. Defining function of lipopolysaccharide O-antigen ligase WaaL using chemoenzymatically synthesized substrates. *J. Biol. Chem.* 2012;287(8):5357-5365. DOI 10.1074/jbc.M111.308486.
- Huerta-Cepas J., Szklarczyk D., Heller D., Hernández-Plaza A., Forslund S.K., Cook H., Mende D.R., Letunic I., Rattei T., Jensen L.J., von Mering C., Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309-D314. DOI 10.1093/nar/gky1085.
- Hug I., Couturier M.R., Rooker M.M., Taylor D.E., Stein M., Feldman M.F. *Helicobacter pylori* lipopolysaccharide is synthesized via a novel pathway with an evolutionary connection to protein N-glycosylation. *PLoS Pathog.* 2010;6(3):e1000819. DOI 10.1371/journal.ppat.1000819.
- Iguchi A., Iyoda S., Kikuchi T., Ogura Y., Katsura K., Ohnishi M., Hayashi T., Thomson N.R. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. *DNA Res.* 2015;22(1):101-107. DOI 10.1093/dnares/dsu043.
- Integrated Taxonomic Information System (ITIS). On-line database. www.itis.gov. (Retrieved 08.21.2022). CC0. https://doi.org/10.5066/F7KH0KKB.
- Jung W.J., Kim S.W., Giri S.S., Kim H.J., Kim S.G., Kang J.W., Kwon J., Lee S.B., Oh W.T., Jun J.W., Park S.C. *Janthinobacterium tractae* sp. nov., isolated from kidney of rainbow trout (*Oncorhynchus mykiss*). *Pathogens.* 2021;10(2):229. DOI 10.3390/pathogens10020229.
- Kalynych S., Morona R., Cygler M. Progress in understanding the assembly process of bacterial O-antigen. *FEMS Microbiol. Rev.* 2014; 38(5):1048-1065. DOI 10.1111/1574-6976.12070.
- Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27-30. DOI 10.1093/nar/28.1.27.
- Kim J.K., Park H.Y., Lee B.L. The symbiotic role of O-antigen of *Burkholderia* symbiont in association with host *Riptortus pedestris*. *Dev. Comp. Immunol.* 2016;60:202-208. DOI 10.1016/j.dci.2016.02.009.
- Koh H.-W., Hur M., Kang M.-S., Ku Y.-B., Ghai R., Park S.-J. Physiological and genomic insights into the lifestyle of arsenite-oxidizing *Herminiimonas arsenitoxidans*. *Sci. Rep.* 2017;7(1):15007. DOI 10.1038/s41598-017-15164-4.
- Lim Y.W., Baik K.S., Han S.K., Kim S.B., Bae K.S. *Burkholderia sordidicola* sp. nov., isolated from the white-rot fungus *Phanerochaete sordida*. *Int. J. Syst. Evol. Microbiol.* 2003;53(5):1631-1636. DOI 10.1099/ijs.0.02456-0.
- Liu B., Furevi A., Perepelov A.V., Guo X., Cao H., Wang Q., Reeves P.R., Knirel Y.A., Wang L., Widmalm G. Structure and genetics of *Escherichia coli* O antigens. *FEMS Microbiol. Rev.* 2020;44(6):655-683. DOI 10.1093/femsre/fuz028.
- Liu B., Knirel Y.A., Feng L., Perepelov A.V., Senchenkova S.N., Wang Q., Reeves P.R., Wang L. Structure and genetics of *Shigella* O antigens. *FEMS Microbiol. Rev.* 2008;32(4):627-653. DOI 10.1111/J.1574-6976.2008.00114.X.
- Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G.A., Sonnhammer E.L.L., Tosatto S.C.E., Paladin L., Raj S., Richardson L.J., Finn R.D., Bateman A. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419. DOI 10.1093/nar/gkaa913.
- Nguyen L.T., Schmidt H.A., Von Haeseler A., Minh B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 2015;32(1):268-274. DOI 10.1093/MOLBEV/MSU300.
- Notredame C., Higgins D.G., Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 2000; 302(1):205-217. DOI 10.1006/JMBI.2000.4042.
- Ofek M., Hadar Y., Minz D. Ecology of root colonizing *Massilia* (Oxalobacteraceae). *PLoS One.* 2012;7(7):e40117. DOI 10.1371/journal.pone.0040117.
- Pereira S.B., Mota R., Vieira C.P., Vieira J., Tamagnini P. Phylum-wide analysis of genes/proteins related to the last steps of assembly and

- export of extracellular polymeric substances (EPS) in cyanobacteria. *Sci. Rep.* 2015;5(1):14835. DOI 10.1038/srep14835.
- Perepelov A.V., Liu B., Senchenkova S.N., Shashkov A.S., Feng L., Wang L., Knirel Y.A. Structure of O-antigen and functional characterization of O-antigen gene cluster of *Salmonella enterica* O47 containing ribitol phosphate and 2-acetimidoylamino-2,6-dideoxy-L-galactose. *Biochemistry (Moscow)*. 2009;74(4):416-420. DOI 10.1134/S0006297909040099.
- Peta V., Raths R., Bücking H. Draft genome sequence of *Massilia* sp. strain ONC3, a novel bacterial species of the Oxalobacteraceae family isolated from garden soil. *Microbiol. Resour. Announc.* 2019; 8(32):e00377-19. DOI 10.1128/MRA.00377-19.
- Ren M., Li X., Zhang Y., Jin Y., Li S., Huang H. *Massilia armeniaca* sp. nov., isolated from desert soil. *Int. J. Syst. Evol. Microbiol.* 2018; 68(7):2319-2324. DOI 10.1099/IJSEM.0.002836.
- Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. Integrative genomics viewer. *Nat. Biotechnol.* 2011;29(1):24-26. DOI 10.1038/nbt.1754.
- Samuel G., Reeves P. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr. Res.* 2003;338(23):2503-2519. DOI 10.1016/j.carres.2003.07.009.
- Sannigrahi S., Arumugasamy S.K., Mathiyarasu J., K.S. Magnetosome-anti-*Salmonella* antibody complex based biosensor for the detection of *Salmonella typhimurium*. *Mater. Sci. Eng. C. Mater. Biol. Appl.* 2020;114:111071. DOI 10.1016/j.msec.2020.111071.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068-2069. DOI 10.1093/bioinformatics/btu153.
- Sheu S.Y., Chou J.H., Bontemps C., Elliott G.N., Gross E., James E.K., Sprent J.I., Young J.P.W., Chen W.M. *Burkholderia symbiotica* sp. nov., isolated from root nodules of *Mimosa* spp. native to north-east Brazil. *Int. J. Syst. Evol. Microbiol.* 2012;62(9):2272-2278. DOI 10.1099/IJS.0.037408-0.
- Sumrall E.T., Röhrig C., Hupfeld M., Selvakumar L., Du J., Dunne M., Schmelcher M., Shen Y., Loessner M.J. Glycotyping and specific separation of *Listeria monocytogenes* with a novel bacteriophage protein tool kit. *Appl. Environ. Microbiol.* 2020;86(13):e00612-20. DOI 10.1128/AEM.00612-20.
- Thakur N., Jain S., Changotra H., Shrivastava R., Kumar Y., Grover N., Vashist J. Molecular characterization of diarrheagenic *Escherichia coli* pathotypes: Association of virulent genes, serogroups, and antibiotic resistance among moderate-to-severe diarrhea patients. *J. Clin. Lab. Anal.* 2018;32(5):e22388. DOI 10.1002/jcla.22388.
- Tuleski T.R., Kimball J., do Amaral F.P., Pereira T.P., Tadra-Sfeir M.Z., de Oliveira Pedrosa F., Maltempi de Souza E., Balint-Kurti P., Monteiro R.A., Stacey G. *Herbaspirillum rubrisubalbicans* as a phytopathogenic model to study the immune system of *Sorghum bicolor*. *Mol. Plant Microbe Interact.* 2020;33(2):235-246. DOI 10.1094/MPMI-06-19-0154-R.
- Valvano M.A. Genetics and biosynthesis of lipopolysaccharide. In: Tang Y.-W., Sussman M., Liu D., Poxton I., Schwartzman J. (Eds.). *Molecular Medical Microbiology*. Academic Press, 2015;55-89. DOI 10.1016/B978-0-12-397169-2.00004-4.
- Wang L., Wang Q., Reeves P.R. The variation of o antigens in gram-negative bacteria. *Subcell. Biochem.* 2010;53:123-152. DOI 10.1007/978-90-481-9078-2_6.
- Xi D., Wang X., Ning K., Liu Q., Jing F., Guo X., Cao B. O-antigen gene clusters of *Plesiomonas shigelloides* serogroups and its application in development of a molecular serotyping scheme. *Front. Microbiol.* 2019;10:741. DOI 10.3389/FMICB.2019.00741.
- Zhang Y.Q., Li W.J., Zhang K.Y., Tian X.P., Jiang Y., Xu L.H., Jiang C.L., Lai R. *Massilia dura* sp. nov., *Massilia albidiflava* sp. nov., *Massilia plicata* sp. nov. and *Massilia lutea* sp. nov., isolated from soils in China. *Int. J. Syst. Evol. Microbiol.* 2006;56(2):459-463. DOI 10.1099/IJS.0.64083-0.

ORCID ID

S.D. Afonnikova orcid.org/0000-0001-7969-8015
A.S. Komissarov orcid.org/0000-0001-6981-7316
P.D. Kuchur orcid.org/0000-0002-9415-577X

Acknowledgements. The authors thank the Collective Center of ICG SB RAS "Bioinformatics" Joint Computational Center for the use of computational resources.

Conflict of interest. The authors proclaim that the research was conducted in the absence of any financial relationship that could be interpreted as a potential conflict of interest.

Received August 29, 2022. Revised October 10, 2022. Accepted October 26, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>


The context signals of mitochondrial miRNAs (mitomiRs) of mammals

O.V. Vishnevsky¹, P.S. Vorozheykin² , I.I. Titov^{1, 2, 3}

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 Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

 pavel.vorozheykin@gmail.com

Abstract. MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression at the post-transcriptional level in the cytoplasm and play an important role in a wide range of biological processes. Recent studies have found that the miRNA sequences are presented not only in the cytoplasm, but also in the mitochondria. These miRNAs (the so-called mitomiRs) may be the sequences of nuclear or mitochondrial origin; some of them are involved in regulation of the mitochondrial gene functions, while the role of others is still unknown. The identification of nucleotide signals, which are unique to mitomiRs, may help to determine this role. We formed a dataset that combined the experimentally discovered mitomiRs in human, rat and mouse. To isolate signals that may be responsible for the mitomiRs' functions or for their translocation from or into mitochondria a context analysis was carried out for the sequences. For three species in the group mitomiRs/non-mitomiRs and the group of all miRNAs from the miRBase database statistically overrepresented 8-letter motifs were identified (p -value < 0.01 with Bonferroni correction for multiple comparisons), for these motifs the patterns of the localization in functionally important regions for different types of miRNAs were found. Also, for the group mitomiRs/non-mitomiRs we found the statistically significant features of the miRNA nucleotide context near the Dicer and Drosha cleavage sites (Pearson's χ^2 test of independence for the first three positions of the miRNA, p -value < 0.05). The observed nucleotide frequencies may indicate a more homogeneous pri-miRNA cleavage by the Drosha complex during the formation of the 5' end of mitomiRs. The obtained results can help to determine the role of the nucleotide signals in the origin, processing, and functions of the mitomiRs.

Key words: miRNA; pre-miRNA; mitomiR; mitochondrion.

For citation: Vishnevsky O.V., Vorozheykin P.S., Titov I.I. The context signals of mitochondrial miRNAs (mitomiRs) of mammals. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):819-825. DOI 10.18699/VJGB-22-99


Контекстные сигналы в митохондриальных микроРНК млекопитающих

О.В. Вишнеvский¹, П.С. Ворожейкин² , И.И. Титов^{1, 2, 3}

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

³ Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

 pavel.vorozheykin@gmail.com

Аннотация. МикроРНК – это малые некодирующие РНК, которые регулируют экспрессию генов на пост-транскрипционном уровне в цитоплазме, и, таким образом, играют важную роль в большом числе биологических процессов. Последние исследования обнаружили присутствие последовательностей микроРНК не только в цитоплазме, но и внутри митохондрий. Такие микроРНК (так называемые митомиры, mitomiRs) могут иметь ядерное или митохондриальное происхождение, при этом для некоторых из них установлена роль в регулировании функций митохондриальных генов, а для большинства она пока неизвестна. Выявление нуклеотидных сигналов, уникальных для митомиров, может помочь определить эту роль. В нашей работе составлена выборка экспериментально обнаруженных митомиров человека, мыши и крысы. С целью выделения сигналов, которые могут быть ответственны за функционирование митомиров и за их транспортировку в митохондрии или из них, осуществлен контекстный анализ для полученных последовательностей митомиров. Для трех видов в группе данных митомиры/не-митомиры и в группе всех микроРНК из базы miRBase выявлены статистически перепредставленные 8-буквенные мотивы (уровень значимости $p < 0.01$ с учетом поправки Бонферрони на множественность сравнения). Для этих мотивов обнаружены закономерности их локализации в функционально значимых участках для разных типов микроРНК. Для рас-

смаатриваемой группы митомир/не-митомир также обнаружены статистически значимые особенности нуклеотидного состава последовательностей микроРНК возле границ разрезания комплексами Drosha/Dicer (критерий независимости χ^2 Пирсона для первых трех позиций микроРНК с уровнем значимости $p < 0.05$). Наблюдаемые частоты нуклеотидов, предположительно, могут указывать на наличие у митомиров (в сравнении с не-митомирами) более однородного разрезания прай-миРНК комплексом Drosha при формировании 5'-конца последовательностей. Результаты работы могут быть полезными для выявления сигналов, принимающих участие в возникновении, процессинге и функциях митомиров.

Ключевые слова: микроРНК; пре-миРНК; митомир; митохондрия.

Introduction

The main pathways of miRNA biogenesis, starting at a cell's nucleus and ending in the cytoplasm, have been studied quite well to date (Bartel, 2018). Studying the nucleotide context of microRNAs and their precursors (pri-/pre-miRNAs) established the presence of signals that can affect the functions of miRNAs as well as their maturation at different stages of biogenesis. The nucleotide sequence of a miRNA can both directly determine its functions and affect the 5'-end cleavage accuracy by Drosha/Dicer complexes, thus forming site-specifically modified miRNAs having a shift in the so-called "seed region", a region from 2 to 7 miRNA nucleotides responsible for its addressing (Starega-Roslan et al., 2015a, b; Rolle et al., 2016).

The presence of motifs in the single-stranded ends (UG; CNNC) and in the basal stem of the pri-miRNAs (CUC/GHG) or in the terminal loop (GU) of the pre-miRNA hairpin can lead to blocking or, conversely, to facilitating miRNA processing (Auyeung et al., 2013; Fang, Bartel, 2015; Nguyen et al., 2015; Starega-Roslan et al., 2015a, b; Rolle et al., 2016; Vorozheykin, Titov, 2020). Apart from the nucleus and cytoplasm, these small RNA sequences, as well as the proteins of their processing complexes, are found in organelles, for example, in mitochondria (Kren et al., 2009; Bandiera et al., 2011; Wang et al., 2015). These observations show there are possibly new pathways for miRNA biogenesis inside a mitochondrion as well as ways for transportation of mature miRNAs between the cytoplasm and mitochondria by yet unknown transport complexes. The existence of such mitochondrial miRNAs (so-called mitomiRs) raises questions about their evolutionary origin and their functions inside and outside organelles and whether they have the structural features enabling their functions and transportation inside or outside mitochondria.

This paper is a review of published materials devoted to experimentally observed miRNAs in a mitochondria. For selected mitomiRs, their sequences' contextual features have been evaluated to investigate the possible influence of nucleotide signals on the origin, processing, and functions of the mitomiRs.

Materials and methods

For our review, miRNA sequences of *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* from the miRBase database (<http://miRBase.org>, edition 22.1) (Kozomara et al., 2019) were selected. The total number of the included sequences comprised 5398.

The information about mitomiRs was obtained from the articles, whose authors experimentally investigated, applying

the RT-qPCR, microarray, qRT-PCR methods, microRNA localization inside and outside the mitochondria of different organisms and tissues (Kren et al., 2009; Bian et al., 2010; Bandiera et al., 2011; Barrey et al., 2011; Mercer et al., 2011; Das et al., 2012; Sripada et al., 2012; Wang et al., 2015). Based on these publications, two sets of sequences were formed for human, mouse, and rat: mitomiRs (652 miRNA sequences observed in mitochondria) and all other miRNAs from the miRBase database (4766 sequences) hereinafter called non-mitomiRs).

To study the features of the sequences of these two groups, a search for statistically overrepresented (p -value < 0.01 with Bonferroni correction for multiple comparisons) oligonucleotide motifs was carried out using the ARGO software (Vishnevsky, Kolchanov, 2005) to perform a *de novo* search for motifs in the 15-letter code for mitomiRs/non-mitomiRs sample pairs and for all miRNAs from the miRBase database. When searching for motifs in the microRNAs from the miRBase database, the software estimated the expected proportion of random sequences with a mononucleotide frequency composition similar to that of an analyzed sample containing a motif for random reasons.

For the obtained motifs, an assessment to estimate their similarity within each of the considered groups and between the two groups was performed. For every pair of motifs, the

Jaccard similarity coefficient was calculated as $\frac{N_{\text{simil}}}{N_{\text{total}}}$, where N_{simil} is the number of all 4-letter nucleotide sequences corresponding to both motifs. The coefficient takes a value from 0 to 1, where 0 indicates a complete difference between the two motifs, and 1 – a complete match.

To estimate the probability of obtaining the Jaccard coefficient for random reasons, the method proposed in (Real, Vargas, 1996) was applied where the random value of the Jaccard coefficient is assumed to be distributed according to the binomial law (up to normalization). For the identified motifs (found in 616 mitomiRs and 4043 non-mitomiRs), an analysis of their localization in miRNA sequence and an analysis of the nucleotide context were performed to identify the heterogeneity of miRNA cleavage from a precursor by the Drosha and Dicer complexes. The localization analysis of all microRNAs from the miRBase database involved the random positions selected within the sequences of an analyzed sample and used as a "contrast" sample.

Results and discussion

In the reviewed publications, 652 unique miRNA identifiers were mentioned. 272 sequences from the found mitomiRs can be characterized as highly reliable, since they were either ad-

ditionally verified by RT-qPCR/qRT-PCR methods, or in the data of microarray experiments, they were observed in greater numbers inside mitochondria than outside them.

It is worth mentioning seven mitomiRs, whose sequences fully present in the human mitochondrial genome: hsa-miR-1974, hsa-miR-1977, hsa-miR-1978, hsa-miR-4461, hsa-miR-4463, hsa-miR-4484, hsa-miR-4485-3p, and that can serve as an additional confirmation of their validity. At the same time, due to miRNA sequence and mitochondrial tRNA imposition, references to the following miRNAs such as hsa-miR-1974, hsa-miR-1977, hsa-miR-1978 were removed from the miRBase database. The hsa-miR-4461 microRNA was also removed from the database since the experimental data obtained for it did not meet the miRNA-annotation requirements. Hence, the sequences that did not correspond to the currently known miRNA biogenesis pathways but could be formed through unknown non-canonical pathways had been excluded from the miRBase database.

For further investigation and comparison of mitomiR characteristics, a sample of non-mitomiRs of a total of 4766 sequences was also used in the study. It includes all human,

mouse, and rat miRNAs from the considered miRBase database, excluding the selected mitomiRs.

Using the ARGO software (Vishnevsky, Kolchanov, 2005), all the two miRNAs groups (mitomiRs/non-mitomiRs group, and all the miRNAs included in the miRBase database) were analyzed. For each of the groups, 40 (Table 1) and 44 (Table 2) 8-nucleotide IUPAC motifs were selected, each having a statistically significant difference in occurrence in miRNA samples in each of the groups ($p < 0.01$ with Bonferroni correction for multiple comparisons). For the motifs within each of the groups, as well as for the motifs from different groups, the Jaccard similarity coefficient average value (averaged over motif pairs excluding zero values) for all three calculations did not exceed 0.02.

For two motifs, KTGCDNDK from the mitomiRs/non-mitomiRs group and KTGABDD from the group of all microRNAs, a maximum Jaccard coefficient of (0.3) and a minimum probability to observe it for random reasons of (0.81) were obtained. These motifs were found in 193 sequences from the mitomiRs/non-mitomiRs group and in 315 sequences in the group of all miRNAs from the miRBase

Table 1. Motifs that have a statistically significant difference in occurrence between the mitomiRs/non-mitomiRs groups

No.	Motif	% of mitomiRs	% of non-mitomiRs	Significance level, p	No.	Motif	% of mitomiRs	% of non-mitomiRs	Significance level, p
1	CAKTSHAN	8.44	0.87	3.8×10^{-26}	21	MWCMBAVH	9.82	2.86	9.7×10^{-8}
2	KTGCANDK	8.90	1.26	5.6×10^{-21}	22	RKTGYWBH	11.81	3.98	1.2×10^{-7}
3	HASHWSBD	28.53	11.89	5.8×10^{-21}	23	GYHSHBDG	18.10	7.97	1.7×10^{-7}
4	WMAGKGCD	6.29	0.54	1.8×10^{-20}	24	YWCMCTBT	5.21	0.85	3.8×10^{-7}
5	MNTVCANK	13.96	3.40	3.6×10^{-20}	25	KKVAACMH	5.98	1.17	8.9×10^{-7}
6	HRVRNTSH	34.97	17.14	1.4×10^{-18}	26	CTNVRBTS	9.66	3.13	2.1×10^{-5}
7	KBAGGTWG	5.21	0.41	8.5×10^{-17}	27	CTRKNBVW	14.88	6.34	2.4×10^{-5}
8	AGSAVCWY	5.21	0.41	8.7×10^{-17}	28	RCABCMHH	6.13	1.40	5.7×10^{-5}
9	RHASHWSB	20.86	7.93	9.1×10^{-16}	29	YCMYWMMM	6.29	1.48	7.5×10^{-5}
10	RCADTSDH	9.97	2.13	7.9×10^{-15}	30	SAGVAMHN	8.13	2.45	2.0×10^{-4}
11	RSTRDRTT	8.13	1.44	4.5×10^{-14}	31	WKMYCMKA	5.21	1.06	2.1×10^{-4}
12	WMDWSCWB	15.49	5.08	1.2×10^{-13}	32	NMYASDGS	10.43	3.79	3.5×10^{-4}
13	HSVYDGDN	44.02	26.58	1.9×10^{-12}	33	KGARNMCY	5.52	1.22	4.0×10^{-4}
14	WRMACWTB	6.13	0.85	3.5×10^{-12}	34	TSRGWSDG	5.98	1.46	9.8×10^{-4}
15	CCHKBWGD	9.36	2.20	2.3×10^{-11}	35	WCCHBTHS	6.60	1.77	1.3×10^{-3}
16	GBYWYWG	12.12	3.63	6.5×10^{-11}	36	SAVWSSCW	6.13	1.59	2.9×10^{-3}
17	KGYWNASW	10.74	3.03	4.8×10^{-10}	37	STRHDGTT	5.06	1.11	3.5×10^{-3}
18	CADKGNTD	8.13	1.79	5.8×10^{-10}	38	NGGCWMDS	7.06	2.07	4.1×10^{-3}
19	GWGSTNVY	9.66	2.60	4.2×10^{-9}	39	HCYBRRCT	5.37	1.26	5.3×10^{-3}
20	WSCAKSWR	6.44	1.24	3.5×10^{-8}	40	YSTSRSTS	5.98	1.55	5.8×10^{-3}

Note. % of mitomiRs is a proportion of mitomiR sequences containing the motif; % of non-mitomiRs is a proportion of non-mitomiR sequences containing the motif.

Here and in the Table 2: The table includes only those motifs whose significance level was $p < 0.01$ with Bonferroni correction for multiple comparisons.

Table 2. Motifs having a statistically significant difference in occurrence between the group of all miRBase base miRNAs and random sequences of similar mononucleotide composition as analyzed miRNAs

No.	Motif	% of miRBase	% of random sequences	Significance level, <i>p</i>	No.	Motif	% of miRBase	% of random sequences	Significance level, <i>p</i>
1	YNCKBYCB	12.09	5.36	1.1×10^{-71}	23	HRHABYRC	5.63	2.99	2.8×10^{-15}
2	BYNCYKYC	11.26	4.94	2.0×10^{-67}	24	GCKSVKBK	6.28	3.46	3.1×10^{-15}
3	AWRYRHWY	6.33	2.10	1.3×10^{-59}	25	KTGYABDD	5.63	3.01	1.0×10^{-14}
4	RHARHRHW	11.79	5.90	1.0×10^{-50}	26	GTWDWHYV	5.15	2.76	9.9×10^{-13}
5	NCKKYCBB	11.09	5.45	1.2×10^{-49}	27	RHBTKTGH	5.94	3.36	1.8×10^{-12}
6	WDYAYDKW	9.21	4.16	2.1×10^{-49}	28	HWYVYAYR	6.11	3.49	3.0×10^{-12}
7	RHAWWYRY	5.06	1.63	4.2×10^{-48}	29	NRRMRSSA	8.40	5.29	4.4×10^{-12}
8	VGGMDVNG	11.92	6.10	4.8×10^{-48}	30	RVKGGMRV	7.88	4.88	5.0×10^{-12}
9	YRTANANV	5.04	1.86	3.9×10^{-37}	31	CBKCYCNV	5.76	3.33	2.2×10^{-10}
10	NHYYVCAG	9.47	4.81	8.4×10^{-37}	32	KCCNBKBC	5.89	3.48	2.0×10^{-9}
11	BTBYCYKY	9.21	4.66	5.7×10^{-36}	33	HATHNYWY	5.70	3.35	3.1×10^{-9}
12	YKHCTYYH	7.97	3.89	2.5×10^{-33}	34	NKGWTDTH	5.06	2.89	9.3×10^{-9}
13	ANBGHWDH	16.11	10.11	4.4×10^{-33}	35	ASDHAVWW	5.37	3.15	2.6×10^{-8}
14	CDGKVNNN	38.28	30.07	7.4×10^{-29}	36	BCDGTKHY	5.30	3.11	5.5×10^{-8}
15	RRMDGNAR	9.63	5.33	5.0×10^{-28}	37	WGDRMHKG	8.99	6.10	1.0×10^{-7}
16	GRGRHDGD	9.10	4.94	6.7×10^{-28}	38	WWWTYRBD	5.41	3.27	7.6×10^{-7}
17	DYAYDGTN	6.02	2.82	6.2×10^{-26}	39	TBTMMYHY	5.30	3.30	5.1×10^{-5}
18	WHAYAHNS	6.24	3.07	2.0×10^{-23}	40	KSRGNBAG	6.31	4.12	6.0×10^{-5}
19	NSDTNTHT	9.10	5.32	1.5×10^{-20}	41	HMCMKYCH	5.44	3.50	6.9×10^{-4}
20	TVYNYVCA	6.39	3.38	1.3×10^{-18}	42	GWSGVDMN	7.88	5.54	1.9×10^{-3}
21	DRYBKTG	5.35	2.72	1.0×10^{-16}	43	GWGHKBAB	5.08	3.26	4.1×10^{-3}
22	TGBRRWKW	5.70	2.98	1.3×10^{-16}	44	TWVTDWRH	5.19	3.37	9.7×10^{-3}

Note. % of miRBase is a proportion of miRBase sequences containing the motif; % of random sequences is an expected by ARGO a proportion of random sequences with a mononucleotide frequency composition similar to analyzed sample containing a motif for random reasons.

database. Apart these two motifs, the Jaccard coefficient for all other considered pairs did not exceed 0.13 with the probability of observing corresponding coefficients for random reasons being less than 0.001, so the observed data showed a low degree of motif coincidence both within each of the groups and between the two groups.

The observed differences in the nucleotide composition between the samples of mitomiRs and non-mitomiRs can act as specific signals for mitomiR processing, e. g., for recognition and transportation of sequences to/from mitochondria by transport complexes or for the implementation of mitomiR specific functions through direct binding to targets in mitochondrial or cellular DNAs. At the same time, the motifs found in the group of all miRNAs may correspond to the signals common for the processing and functioning of miRNAs, regardless of their localization.

For both considered groups the first motif position tended to be located at the beginning of a miRNA (Fig. 1), so the

maximum proportion of sequences was observed for the motifs with their start being at positions 1–3. For the obtained observations, a statistically significant dependence of a miRNA-sequence type on the positions of a motif start (Pearson’s χ^2 test of independence, *p*-values 4.46×10^{-2} and 6.58×10^{-5} for the mitomiRs/non-mitomiRs group and the miRbase miRNAs/random positions in miRNAs group, respectively). At the same time, for the mitomiRs, in contrast to the other samples, a significant reduction in the number of miRNAs whose motif start was located at positions 8–10 of a microRNA was observed. A possible reason for that was that the so-called seed region of all miRNAs (both mitomiRs and non-mitomiRs) is the most conservative and significant region in terms of its functionality, and therefore the considered 8-letter conservative motifs often take this region.

In contrast to this, the motifs starting from 8 to 10 position in a miRNA are often localized in the region of the so-called additional seed (~13–16), which is supposedly less conservative

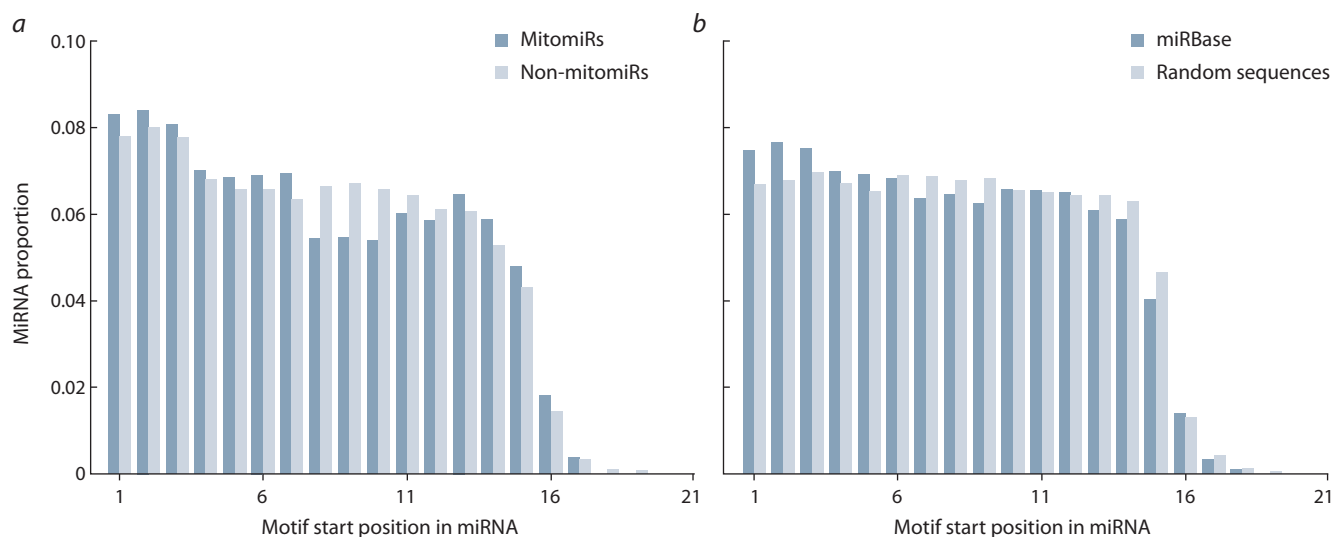


Fig. 1. MiRNAs proportion depending on the motif-start positions found in the sequences of mitomiRs and non-mitomiRs (a) and in the miRNAs from the miRBase database (b).

For each miRNA with a motif, its position starting from the 5' end of the sequence was determined. If one motif occurred several times in a microRNA or several motifs occurred once in a microRNA, each occurrence was considered independently and generated a data structure (microRNA, motif position). The graph was normalized for the total number of structures obtained for all motifs. The decrease in observations in the positions whose numbers were greater than 15 was due to the variability in the lengths of miRNA sequences changing from 15 to 28 nucleotides.

in mitomiRs and less often participates in microRNA binding to the target if compared to non-mitomiRs (see Fig. 1, a). However, since the sample of all miRNAs mostly consisted of non-mitomiRs, its observation results approximately coincided with those obtained for non-mitomiRs.

For the detected motifs, different localization patterns within a miRNA sequence were observed. One motif could be observed as in several different miRNAs with different localizations within the sequences (e.g., the KTGANDK motif with a significance level of $p = 5.6 \times 10^{-21}$ started at position 14 from the 5' end of the hsa-miR-92a-1-5p mitomiR, and at position 2 from the 5' end of the mmu-miR-19b-3p mitomiR) as within one miRNA, including cases that did not intersect each other (e.g., in the hsa-miR-33a-5p mitomiR, the KTGANDK motif occurs twice, starting from position 1 and from position 12 from the 5' end).

Hence, the variability of motif localization in a miRNAs may indicate both the functional importance of these nucleotide signals for these miRNAs and the possible involvement of the signals in microRNA processing, in particular, in the selection and transportation of mitomiR sequences between a mitochondrion and cytoplasm.

For the considered miRNA groups, an increase in the proportion of the sequences where motifs began in positions 1–3 was observed (see Fig. 1), so sequence analysis of this region, but only for those mitomiR and non-mitomiR sequences in which motifs has previously been found, was performed. For the first three positions of the 5' end of 5p- and 3p-miRNAs, the positional frequencies of nucleotide occurrence were calculated (Fig. 2).

In the mitomiRs from a pre-miRNA's 5p-branch, U was predominantly observed in the first position and G was very rarely found, while A or G were mainly detected in the second position (see Fig. 2, a). In non-mitomiRs, an increase in

the first position of the number of G and A nucleotides and a decrease in the number of U were observed (see Fig. 2, b). For the Drosha cleavage site, an inversion in 2–3 positions between G in non-mitomiRs and A in mitomiRs was detected. For each of the three positions, the nucleotide frequencies showed dependence on a miRNA type (Pearson's χ^2 test of independence, p -values 2.89×10^{-31} , 1.03×10^{-28} , and 1.79×10^{-42} for the 1, 2 and 3 positions, respectively), while the third position demonstrated the most significant difference in frequencies between miRNA types, in contrast to the first and second ones.

Comparing the observed nucleotide context of the 5' ends for mitomiRs/non-mitomiRs with the results of a study that investigated pre-miRNA cleavage accuracy by Drosha and Dicer complexes (Starega-Roslan et al., 2015b), it can be assumed that the Drosha cleavage was more accurate for the mitomiRs from the 5p-branch of pre-miRNAs than for non-mitomiRs, in other words, a more homogeneous 5' end and a corresponding seed region were formed for mitomiRs, which may be the evidence of greater conservatism of mitomiR functions in comparison to those of non-mitomiRs. The detected signals of mitomiR cleavage homogeneity could be an indication either of the possible existence of a more accurate Drosha-like complex for miRNA processing in mitochondria or of the possible compensation of inaccurate cleavage by the nucleotide composition of the pri-miRNA sequences selected for processing by the Drosha complex.

For the 5p-non-mitomiRs, the context shifted towards heterogeneous cleavage, i.e., more active site-specific miRNA modification, and in this case, the non-mitomiRs could act as a functional-variability factor. It can be assumed that mitochondria do not "tolerate" the variability of "their" miRNAs and eliminated regulatory-sequence isoforms in the course of evolution, so the observed mitomiRs may be the remaining

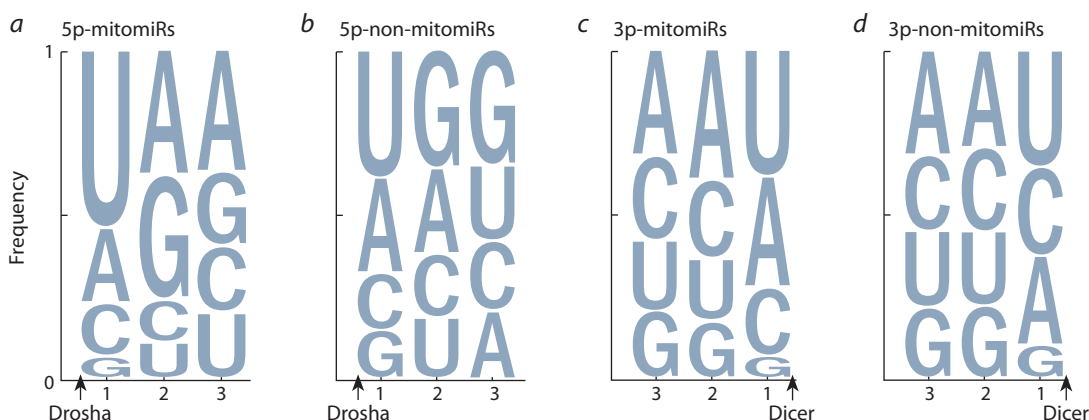


Fig. 2. Nucleotide occurrence frequency for the first three positions starting from the 5' end of the microRNA from 5p- and 3p-branches of pre-miRNA for mitomiRs (a, c) and non-mitomiRs (b, d) samples in the sequences where motifs were found. The sizes of the letters are proportional to the frequencies. The X-axis displays the position numbers in microRNA, starting from the 5' end. The arrows show the Dicer or Drosha cleavage sites. For positions 1–3 in a 5p-miRNA and for positions 1–2 in a 3p-miRNA presented a statistically significant dependence of miRNA sequence type on nucleotide occurrence frequency for a considered position (Pearson's χ^2 test of independence, p -value: 2.89×10^{-31} , 1.03×10^{-28} , 1.79×10^{-42} , 1.17×10^{-9} , 3.23×10^{-10} for respective positions).

conservative sequences that originated in the times mitochondrial ancestors were domesticated.

For mitomiRs and non-mitomiRs from the 3p branch of pre-miRNA, no noticeable differences in the nucleotide context were observed, except for the inversion in the first position of the second and third most popular nucleotides (see Fig. 2, c and d). Statistically significant dependence of positional nucleotide frequencies on a miRNA type was demonstrated only at the first and second positions (Pearson's χ^2 test of independence, p -values: 1.17×10^{-9} and 3.23×10^{-10} , respectively). Comparison of the observed nucleotide frequencies against the results obtained by (Starega-Roslan et al., 2015b) did not allow us to make unambiguous conclusions about the cleavage quality of the 5' end of the 3p-miRNA by the Dicer complex.

Conclusion

In the present study, a sample of experimentally confirmed mitomiRs was formed and a nucleotide analysis of their sequences was performed. For the mitomiRs/non-mitomiRs group and the group of all microRNAs from the miRbase database, statistically overrepresented 8-letter IUPAC motifs within miRNA sequences were found. These motifs demonstrated that mitomiR sequences may represent a new, non-canonical class of miRNAs. While the motifs for the mitomiRs/non-mitomiRs group could act as signals for mitomiR processing, e. g., participating in mitomiR transportation to/from mitochondria or for mitomiR function implementation through binding to targets in mitochondrial or cellular DNAs, the motifs of the group of all microRNAs could correspond to the signals common for the processing and functions of miRNAs, regardless of their localization in the cell.

The nucleotide context of the mitomiRs (if compared to that of the non-mitomiRs) near the 5' end formed by Drosha/Dicer cleavage could presumably indicate a more uniform formation of the 5' end of mitomiR sequences and, thus, a more conserved functionality of these sequences.

References

- Auyeung V.C., Ulitsky I., McGeary S.E., Bartel D.P. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*. 2013;152(4):844-858. DOI 10.1016/j.cell.2013.01.031.
- Bandiera S., Rüberg S., Girard M., Cagnard N., Hanein S., Chrétien D., Munnich A., Lyonnet S., Henrion-Caude A. Nuclear outsourcing of RNA interference components to human mitochondria. *PLoS One*. 2011;6(6):e20746. DOI 10.1371/journal.pone.0020746.
- Barrey E., Saint-Auret G., Bonnamy B., Damas D., Boyer O., Gidrol X. Pre-microRNA and mature microRNA in human mitochondria. *PLoS One*. 2011;6(5):e20220. DOI 10.1371/journal.pone.0020220.
- Bartel D.P. Metazoan microRNAs. *Cell*. 2018;173(1):20-51. DOI 10.1016/j.cell.2018.03.006.
- Bian Z., Li L.-M., Tang R., Hou D.-X., Chen X., Zhang C.-Y., Zen K. Identification of mouse liver mitochondria-associated miRNAs and their potential biological functions. *Cell Res*. 2010;20(9):1076-1078. DOI 10.1038/cr.2010.119.
- Das S., Ferlito M., Kent O.A., Fox-Talbot K., Wang R., Liu D., Raghavachari N., Yang Y., Wheelan S.J., Murphy E., Steenbergen C. Nuclear miRNA regulates the mitochondrial genome in the heart. *Circ. Res*. 2012;110(12):1596-1603. DOI 10.1161/CIRCRESAHA.112.267732.
- Fang W., Bartel D.P. The menu of features that define primary microRNAs and enable *de novo* design of microRNA genes. *Mol. Cell*. 2015;60(1):131-145. DOI 10.1016/j.molcel.2015.08.015.
- Kozomara A., Birgaoanu M., Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019;47(D1):D155-D162. DOI 10.1093/nar/gky1141.
- Kren B.T., Wong P.Y.-P., Sarver A., Zhang X., Zeng Y., Steer C.J. MicroRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis. *RNA Biol*. 2009;6(1):65-72. DOI 10.4161/rna.6.1.7534.
- Mercer T.R., Neph S., Dinger M.E., Crawford J., Smith M.A., Shearwood A.-M.J., Haugen E., Bracken C.P., Rackham O., Stamatoyannopoulos J.A., Filipovska A., Mattick J.S. The human mitochondrial transcriptome. *Cell*. 2011;146(4):645-658. DOI 10.1016/j.cell.2011.06.051.
- Nguyen T.A., Jo M.H., Choi Y.-G., Park J., Kwon S.C., Hohng S., Kim V.N., Woo J.-S. Functional anatomy of the human microprocessor. *Cell*. 2015;161(6):1374-1387. DOI 10.1016/j.cell.2015.05.010.

- Real R., Vargas J. M. The probabilistic basis of Jaccard's index of similarity. *Syst. Biol.* 1996;45(3):380-385. DOI 10.1093/sysbio/45.3.380.
- Rolle K., Piwecka M., Belter A., Wawrzyniak D., Jeleniewicz J., Barciszewska M.Z., Barciszewski J. The sequence and structure determine the function of mature human miRNAs. *PLoS One.* 2016; 11(3):e0151246. DOI 10.1371/journal.pone.0151246.
- Sripada L., Tomar D., Prajapati P., Singh R., Singh A.K., Singh R. Systematic analysis of small RNAs associated with human mitochondria by deep sequencing: detailed analysis of mitochondrial associated miRNA. *PLoS One.* 2012;7(9):e44873. DOI 10.1371/journal.pone.0044873.
- Starega-Roslan J., Galka-Marciniak P., Krzyzosiak W.J. Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer. *Nucleic Acids Res.* 2015a;43(22):10939-10951. DOI 10.1093/nar/gkv968.
- Starega-Roslan J., Witkos T., Galka-Marciniak P., Krzyzosiak W. Sequence features of Drosha and Dicer cleavage sites affect the complexity of isomiRs. *Int. J. Mol. Sci.* 2015b;16(12):8110-8127. DOI 10.3390/ijms16048110.
- Vishnevsky O.V., Kolchanov N.A. ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucleic Acids Res.* 2005;33(Web Server Iss.):W417-W422. DOI 10.1093/nar/gki459.
- Vorozheykin P.S., Titov I.I. Erratum to: How animal miRNAs structure influences their biogenesis. *Russ. J. Genet.* 2020;56(8):1012-1024. DOI 10.1134/S1022795420220019.
- Wang W.-X., Visavadiya N.P., Pandya J.D., Nelson P.T., Sullivan P.G., Springer J.E. Mitochondria-associated microRNAs in rat hippocampus following traumatic brain injury. *Exp. Neurol.* 2015;265:84-93. DOI 10.1016/j.expneurol.2014.12.018.

ORCID ID

I.I. Titov orcid.org/0000-0002-2691-3292

Acknowledgements. The OV and IT works were supported by the Russian State Budgetary Project No. FWNR-2022-0020.

The IT work was supported by the Kurchatov Genomic Center of ICG SB RAS (075-15-2019-1662).

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

Received September 7, 2022. Revised November 17, 2022. Accepted November 17, 2022.

Original Russian text <https://sites.icgbio.ru/vogis/>


Small world of the miRNA science drives its publication dynamics

A.B. Firsov¹ , I.I. Titov^{2, 3}

¹ A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

³ Novosibirsk State University, Novosibirsk, Russia

 artyomfirsov@mail.ru

Abstract. Many scientific articles became available in the digital form which allows for querying articles data, and specifically the automated metadata gathering, which includes the affiliation data. This in turn can be used in the quantitative characterization of the scientific field, such as organizations identification, and analysis of the co-authorship graph of those organizations to extract the underlying structure of science. In our work, we focus on the miRNA science field, building the organization co-authorship network to provide the higher-level analysis of scientific community evolution rather than analyzing author-level characteristics. To tackle the problem of the institution name writing variability, we proposed the k-mer/n-gram boolean feature vector sorting algorithm, KOFER in short. This approach utilizes the fact that the contents of the affiliation are rather consistent for the same organization, and to account for writing errors and other organization name variations within the affiliation metadata field, it converts the organization mention within the affiliation to the K-Mer (n-gram) Boolean presence vector. Those vectors for all affiliations in the dataset are further lexicographically sorted, forming groups of organization mentions. With that approach, we clustered the miRNA field affiliation dataset and extracted unique organization names, which allowed us to build the co-authorship graph on the organization level. Using this graph, we show that the growth of the miRNA field is governed by the small-world architecture of the scientific institution network and experiences power-law growth with exponent 2.64 ± 0.23 for organization number, in accordance with network diameter, proposing the growth model for emerging scientific fields. The first miRNA publication rate of an organization interacting with already publishing organization is estimated as $0.184 \pm 0.002 \text{ year}^{-1}$.

Key words: k-mer; n-gram; miRNA; digital library; organization co-authorship; small world.

For citation: Firsov A.B., Titov I.I. Small world of the miRNA science drives its publication dynamics. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2022;26(8):826-829. DOI 10.18699/VJGB-22-100

Свойства малого мира научных организаций определяют динамику публикационной активности в области мирНК

А.Б. Фирсов¹ , И.И. Титов^{2, 3}

¹ Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия

² Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

³ Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

 artyomfirsov@mail.ru

Аннотация. Многие научные статьи стали доступны в цифровом виде, что позволяет запрашивать данные статей и, в частности, автоматически собирать метаданные, включая данные об аффилиации. Это, в свою очередь, можно использовать для количественных оценок научной области, например для идентификации организаций и анализа графа соавторства этих организаций для извлечения базовой структуры науки. В настоящей работе рассмотрена область исследования микроРНК, а именно граф соавторства организаций и анализ его эволюции. Чтобы решить проблему вариативности написания названия организаций, был предложен алгоритм сортировки логических векторов признаков k-mer/n-gram. В нем используется тот факт, что содержание аффилиации довольно консистентно для одной и той же организации. Для учета ошибок написания и других артефактов названия организации в поле метаданных аффилиации наш подход преобразует упоминание организации внутри аффилиации в K-Mer (n-gram) булевый вектор присутствия. Далее

векторы всех аффилиаций из набора данных лексикографически сортируются, образуя группы упоминаемых организаций. Таким подходом был кластеризован набор данных аффилиаций в области исследования микроРНК и определены названия уникальных организаций, что позволило построить граф соавторства на уровне научных организаций. С помощью этого графа показано, что рост области исследования микроРНК контролируется архитектурой малого мира сети научных организаций и испытывает степенной рост с показателем степени 2.64 ± 0.23 для числа организаций в соответствии с диаметром сети, предлагая модель роста новых научных направлений. Скорость публикации первой статьи по микроРНК у организации при ее взаимодействии с другой организацией, уже публиковавшейся в этой области, аппроксимируется как $0.184 \pm 0.002 \text{ год}^{-1}$.

Ключевые слова: k-mer; n-gram; микроРНК; электронная библиотека; соавторство организаций; малый мир.

Introduction

Scientific structures stimulate the productivity of scientific work by providing researchers with material and technical conditions and a scientific environment. One of the factors for the effectiveness of scientific work is the interaction of researchers in the form of an exchange of ideas or joint work and is manifested in the form of scientific publications co-authorship. Analysis of the co-authorship of research institutions, rather than characteristics at the authors level, makes it possible to provide a higher-level analysis of the evolution of the scientific community, in particular the organization of “invisible colleges” or the development of international cooperation on a global scale (Leydesdorff et al., 2013). Such studies are aimed at finding the reasons for competition and cooperation in specific areas of research (Wagner, Leydesdorff, 2005), as well as identifying patterns of international publication activity (Ribeiro et al., 2017). In general, in order to understand the structure of the scientific community and the process of knowledge spreading in the field of science, analysis should be carried out both at the author level and at the organization level.

A graph is a small world if $L \propto \log(N)$, where L is the average shortest distance of the graph, N is the number of graph vertices. In other words, any two vertices are reachable from the other through a small number of hops through other vertices, but the probability that they are adjacent is small.

This type of networks are found in many real-world phenomena, such as the spread of the infection (Liu et al., 2015), neural connections (Muldoon et al., 2016), etc. The analysis of the effect of the small world in the knowledge spreading (Shi, Guan, 2016) is of particular interest, and therefore our study aims to check whether the interaction graph of organizations in the miRNA research field is a small world.

Since in a small world the vertices are reachable between each other via a small number of hops, processes such as the spread of the infection or knowledge must occur differently than in a regular graph.

To determine that a graph is a small world, various criteria have been proposed in several works (Watts, Strogatz, 1998; Newman et al., 2000). In our work, we chose a categorical criterion to identify the small world effect in a network of microRNA organizations co-authorship, following (Humphries, Gurney, 2008), where the authors introduced a measure of the “small-world-ness”:

$$S = \frac{CC_G}{CC_{\text{rand}}} \cdot \frac{L_G}{L_{\text{rand}}}.$$

In the equation above, CC_G is the clustering coefficient of graph G , L_G is the average length of the shortest paths of graph G , CC_{rand} and L_{rand} are the parameters of a random graph with random uniform edge placement with the same number of nodes and edges as graph G .

The knowledge spreading process can be interpreted as a process of “information contagion” where, through an intermediate host (scientific publications), organizations can be inspired by a particular area of research and start publishing articles themselves. Such a process can be modeled using the Susceptible, Infectious, Recovered (SIR) model (Goffman, Newell, 1964). Within the framework of this model, a system of differential equations is compiled that simulates the dynamics of infection and recovery of subjects. In the simplest case of a homogeneous environment, the solution to these equations at short times is the exponential growth in the number of infected subjects.

In (Vazquez, 2006), the author models the incidence rate using the SIR model for problems where transmission graphs are known and have the small world property (Muldoon et al., 2016). The author adapts the SIR propagation model to a spanning tree (AST) representation of the original graph and obtains the exact normalized incidence rate for the AST, $\rho(t)$, which approximates this rate for the original graph. Thus, given that the graph has the small world property, there is an exact solution to the normalized infection rate for the AST, which is the approximation for the original graph:

$$\rho(t) = \lambda \frac{(\lambda t)^{D-1}}{(D-1)!} e^{-(\lambda+\mu)t} \left[1 + O\left(\frac{t_0}{t}\right) \right],$$

where λ and μ are, respectively, the rates of infection and recovery within the framework of the SIR model, D is the average shortest distance of the graph, t_0 is the transition time between modes. The graph, in addition to having the characteristics of a small world, must satisfy one of the conditions for γ (the exponent of the power law distribution of degrees of vertices) and ν (the Pearson correlation coefficient of the degree between pairs of connected nodes) (Vazquez, 2006):

$$\begin{aligned} \gamma &> 3, \quad \nu > 0, \\ 2 \leq \gamma \leq 3, \quad \nu > -1, \quad 3 - \gamma + \nu > 0. \end{aligned}$$

Methods and materials

The PubMed digital library was used to collect the miRNA research area affiliation dataset. From these affiliations, mentions of the organizations were extracted. To do this, a key-

word-based approach was used to identify which part of the affiliation contains what information about the mention of the organization (organization name, country, city, etc.).

An example of splitting an affiliation into mentions of organizations with a country identification for an article with PubMed ID 19996210

(1) Authors' Affiliations: Cancer Genetics, Kolling Institute of Medical Research; Department of Endocrinology; Department of Anatomical Pathology, Royal North Shore Hospital, St. Leonards, New South Wales, Australia; Department of Surgery, Bankstown Hospital, Bankstown, New South Wales, Australia; South Western Sydney Clinical School, University of New South Wales; Endocrine Surgical Unit, University of Sydney; Department of Surgery, Liverpool Hospital, Sydney, New South Wales, Australia; Endocrine Surgical Unit, University of California Los Angeles; and Division of Hematology and Oncology, Department of Medicine, University of California Los Angeles School of Medicine, Los Angeles, California.	1. kolling institute of medical research, Australia 2. royal north shore hospital, Australia 3. bankstown hospital, Australia 4. university of new south wales, Australia 5. university of sydney, Australia 6. liverpool hospital, Australia 7. university of california los angeles, UNKNOWN 8. university of california los angeles, school of medicine, UNKNOWN
--	--

Then, for all these mentions, a dictionary of unique K-Mers (n-grams) was built, where $K = 2$, and for each mention, a Boolean vector of the presence of a certain K-Mer in this mention was formed. Next, these mention vectors were sorted

lexicographically to obtain a list of vectors, in which similar mentions are grouped by design. After that, for each adjacent pair of mentions, the distance according to the Dice metric was calculated, and if it exceeded the specified threshold, this was the evidence that the mentions belong to different clusters, which gives us a grouping of mentions (see the Table).

These grouped mentions contain references to the same organization; so, in the next step, we can build an organization co-authorship graph by identifying which organizations published the same article together.

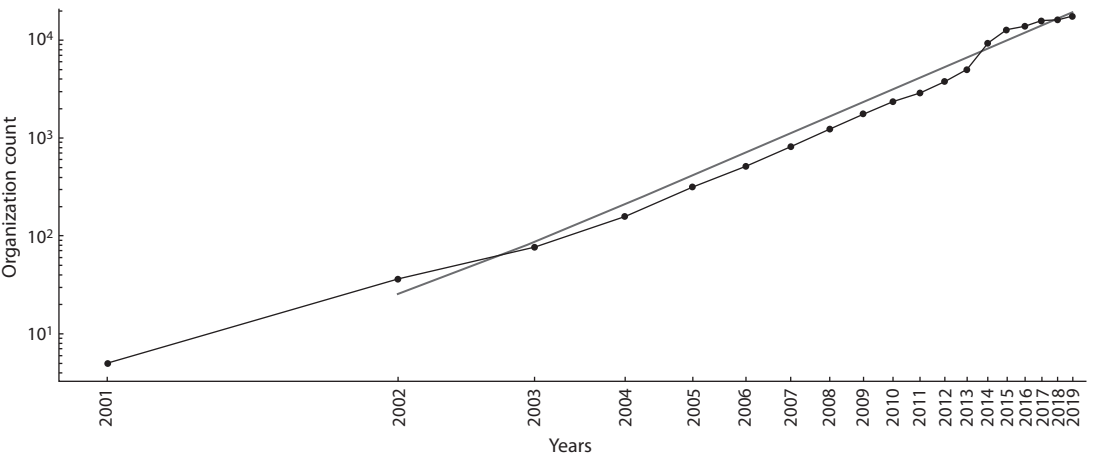
Results

The analysis of the structural characteristics of the graph of scientific organizations in the miRNA research field shows that this graph satisfies the criteria of a small world (Muldoon et al., 2016) with the exponent of the degree of power distribution $\gamma = 2.01$ and the assortativity coefficient of the degrees of graph vertices $v = -0.03$. Therefore, for the number of scientific organizations with publications in the field, one can expect a power-law growth according to the model (Vazquez, 2006). The model (Vazquez, 2006) states that the initial growth in the number of vertices has a power-law dependence with the exponent $D - 1$, where D is the average length of the shortest paths in the graph. For the graph of scientific organizations of the microRNA research field $D = 3.46$, and the approximated power parameter $D - 1 = 2.64 \pm 0.23$ (see the Figure), which gives a deviation of about 7 % from what is predicted by the model.

An example of organizations identification

#	Mention	2-Mer Boolean vector	Dice metric
1	institute	1111111100000000	0.2
2	insitute	1111100100001000	0.429
3	institutue	1111011000010000	0.834
4	center	0000100011100100	0.4
5	centre	0000000011100011	

Note. The threshold value is 0.8, $K = 2$. The distance between elements 3, 4 exceeds the threshold value, which leads to the division of elements into clusters. 2-Mer examples – in, ns, st, ti, it, tu, ...



Annual number of organizations that published an article in the field of the microRNA research as a function of time in double logarithmic coordinates.

Approximation of the “information contagion” rate gives the rate $\lambda = 0.184 \pm 0.002 \text{ year}^{-1}$, which characterizes the rate of the first microRNA publication by an organization in co-authorship with another organization that already published in this field.

Analysis of the subgraph of Russian scientific institutions in the miRNA research field shows that the activity of Russian organizations is inferior to the average activity of organizations in the field (the average number of publications per organization is 0.92 in Russia against 21.5 on average in the field). At the same time, the Russian community turns out to be denser: the clustering coefficient of the subgraph of Russian organizations exceeds the average for the field with the value of 0.708 for Russian organizations compared to the 0.361 for the microRNA field average. The US is Russia’s most active partner in international cooperation with 50 joint publications. However, US-Russian cooperation is unstable and decentralized, and the leaders in active cooperation with Russian organizations are the German Center for Cancer Research, Harbin Medical University, and Karolinska Institute (6 joint publications each).

Discussion

Understanding the productivity factors of research organizations and the dynamics of their publication activity is important for science management. In addition to algorithms for automatic identification of organizations, projects such as ror.org are actively developing, and are aimed at identifying scientific institutions by assigning unique identifiers to them (similar to orcid.org for authors). These projects simplify the identification of organizations but require the acceptance of the use of such projects by the authors of publications, since in order to be able to fully identify each organization, it is necessary to indicate the ror.org identifier for each affiliation from the publication, which cannot currently be guaranteed. Therefore, in the near future, automatic identification algorithms for organizations will stay relevant.

In our work, the data presented was gathered as of 2019, and at the current moment the structure of the graph could change. In addition, the data in the PubMed library can be updated retrospectively. Nevertheless, data from publications as of January 23, 2022 show that the picture of the evolution of the miRNA field has not fundamentally changed (data not shown). The new geopolitical reality will inevitably affect the structure of interaction and co-authorship in scientific fields. However, due to the time delay in the visible results of cooperation, a change in scientific cooperation will not appear in the databases until 2024.

Conclusion

One of the models of the development of new knowledge areas is the “information contagion” model, in which new ideas are randomly distributed among researchers, infecting more and more of them (Goffman, Newill, 1964). The distribution law can be determined by the structure of the environment. In this work, it was shown that the organization co-authorship graph in the microRNA research field is a small world and, as a result, the publication activity of the area demonstrates a power-law growth according to the model (Vazquez, 2006). The slower than exponential growth occurs due to the “self-avoidance” of propagation paths in compact networks of the small world: when the next node of the small world is “infected” with information, there is a high probability that this node has already been “infected” by an alternative path. The co-authorship graph for our analysis was built using the organization mention clustering algorithm based on sorting K-Mer boolean feature vectors (KOFER).

References

- Goffman W., Newill V.A. Generalization of epidemic theory. An application to the transmission of ideas. *Nature*. 1964;204(4955):225-228. DOI 10.1038/204225a0.
- Humphries M.D., Gurney K. Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. *PLoS One*. 2008;3(4):e0002051. DOI 10.1371/journal.pone.0002051.
- Leydesdorff L., Wagner C., Park H., Adams J. International collaboration in science: the global map and the network. *Prof. Inf.* 2013; 22(1):1-18. DOI 10.3145/epi.2013.ene.12.
- Liu M., Li D., Qin P., Liu C., Wang H., Wang F. Epidemics in interconnected small-world networks. *PLoS One*. 2015;10(3):e0120701. DOI 10.1371/journal.pone.0120701.
- Muldoon S., Bridgeford E., Bassett D. Small-world propensity and weighted brain networks. *Sci. Rep.* 2016;6:22057. DOI 10.1038/srep22057.
- Newman M.E.J., Moore C., Watts D.J. Mean-field solution of the small-world network model. *Phys. Rev. Lett.* 2000;84(14):3201-3204. DOI 10.1103/PhysRevLett.84.3201.
- Ribeiro L., Rapini M., Silva L., Albuquerque E.M. Growth patterns of the network of international collaboration in science. *Scientometrics*. 2018;114:159-179. DOI 10.1007/s11192-017-2573-x.
- Shi Y., Guan J. Small-world network effects on innovation: evidences from nanotechnology patenting. *J. Nanopart. Res.* 2016;18:329. DOI 10.1007/s11051-016-3637-1.
- Vazquez A. Spreading dynamics on small-world networks with connectivity fluctuations and correlations. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* 2006;74:056101. DOI 10.1103/PhysRevE.74.056101.
- Wagner C., Leydesdorff L. Network structure, self-organization and the growth of international collaboration in science. *Res. Policy*. 2005; 34(10):1608-1618. DOI 10.1016/j.respol.2005.08.002.
- Watts D.J., Strogatz S.H. Collective dynamics of ‘small-world’ networks. *Nature*. 1998;393(6684):440-442. DOI 10.1038/30918.

ORCID ID

A. Firsov orcid.org/0000-0002-7681-1032
I.I. Titov orcid.org/0000-0002-2691-3292

Acknowledgements. The work of IT was supported by the Russian State Budgetary Project FWN-2022-0020.

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

Received September 7, 2022. Revised November 10, 2022. Accepted November 10, 2022.

Алфавитный указатель авторов статей, опубликованных в журнале в 2022 г.

Агеева Е.В. 7, 675
Адамовская А.В. 8, 733
Айдаров А.Н. 5, 413
Акиншин А.А. 8, 758
Аклеев А.В. 1, 50
АлБосале А.Х. 1, 59
Алексеев Я.И. 6, 544
Аль-Накиб Е.А. 7, 645
Амстиславская Т.Г. 4, 365
Амстиславский С.Я. 4, 365; 5, 431
Андреева Л.В. 6, 537
Анисимова Е.А. 6, 560
Антонов А.А. 4, 349
Антонова О.Ю. 2, 128
Антропова Е.А. 8, 733
Ануфриев К.Э. 8, 743
Атопкин Д.М. 3, 290
Аульченко Ю.С. 4, 378
Афанасенко О.С. 3, 272
Афонников Д.А. 8, 787
Афонникова С.Д. 8, 810
Аюпова Н.Б. 8, 758

Багиров В.А. 4, 378
Бажан Н.М. 2, 159
Бакулина А.Ю. 3, 240
Балабова Д.В. 3, 240
Балыбина Н.Ю. 2, 159
Баранов К.О. 2, 159
Баттулин Н.Р. 4, 402
Белавин П.А. 3, 327
Белан И.А. 7, 662
Беленькая С.В. 3, 240
Белкасем эль-Амрани 5, 442
Белов А.Н. 3, 240
Беловежец Л.А. 6, 568
Белчев И. 2, 139
Белькова Н.Л. 5, 495
Берман Д.И. 1, 109
Блинова Е.А. 1, 50
Блинова С.А. 6, 544
Богданова В.С. 4, 359
Богомоллов А.Г. 1, 96; 8, 798
Богословская Т.Ю. 3, 319
Бойко А.П. 7, 599
Бондарь А.А. 3, 240
Бородин П.М. 4, 378
Бочарникова М.Е. 8, 787
Брусенцев Е.Ю. 4, 365; 5, 431
Брызгалов Л.О. 1, 65
Букин Ю.С. 1, 74, 86
Булахова Н.А. 1, 109
Бурменко Ю.В. 7, 622
Бурыгин Г.Л. 5, 422
Быков Р.А. 6, 553

Васильев В.Б. 3, 319
Васильева О.Ю. 1, 22
Вензель А.С. 8, 733
Вергунов Е.Г. 8, 765
Весна Э. 8, 806
Вишневский О.В. 8, 798, 819
Вишнякова М.А. 7, 599
Водясова Е.А. 3, 290
Волкова Н.А. 4, 378
Воловик В.Т. 4, 349
Воробьева С.С. 5, 467
Ворожейкин П.С. 8, 819
Воронина А.В. 2, 146
Воронина Е.Н. 2, 188
Воропаева Н.М. 5, 495

Гавриленко А.Д. 8, 733
Гаврилова Е.В. 4, 394
Газизова Г.Р. 1, 109
Герасимова С.В. 2, 153
Герасимчук А.Л. 5, 449
Гисматулина Ю.А. 1, 22
Глушаков Д.А. 7, 609
Голохваст К.С. 7, 637
Голубев С.Н. 5, 477
Голубятников В.П. 8, 758
Гончаров Н.П. 7, 662
Горбунова М.Е. 6, 560
Григорова Е.В. 5, 495
Грин И.Р. 4, 341
Гультяева Е.И. 6, 537
Гурина В.В. 6, 568

Давыдова Ю.Д. 2, 179
Дамаров И.С. 1, 65
Деятериков А.П. 8, 780
Дейнеко Е.В. 3, 327
Деменков П.С. 8, 733
Джос Е.А. 7, 652
Дмитриева Е.В. 3, 290
Додонова Е.А. 6, 560
Дорогина О.В. 1, 22
Драгов Р.Г. 6, 515
Драчкова И.А. 3, 227
Дружин А.Е. 6, 537
Дубинина А.Д. 2, 159
Душкин В.А. 4, 349
Дьяченко Е.А. 7, 652
Дюдеева Е.С. 1, 5

Евдокименко С.Н. 7, 622
Евдокимов М.Г. 7, 609
Евсеева Н.В. 5, 422
Егорова А.А. 2, 153
Елаткин Н.П. 3, 298

Елизарова И.А. 6, 560
Елизарова С.А. 6, 583
Ельчанинов В.В. 3, 240
Еникеева Р.Ф. 2, 179
Ермаков М.С. 1, 14
Ефимов В.М. 7, 662
Ефремов Г.И. 6, 507

Жарков Д.О. 4, 341
Жданова И.Н. 5, 486
Жемчужина Н.С. 6, 583

Зайнуллин Л.И. 6, 560
Захаренко А.М. 7, 637
Захарова Ф.М. 3, 319
Зеленских М.О. 8, 773
Землянская Е.В. 8, 721, 798
Злобин А.С. 4, 378
Золотарева К.А. 8, 798
Зуев Е.В. 7, 662
Зуева Г.А. 1, 22

Иванисенко В.А. 2, 121; 8, 733
Иванисенко Н.В. 8, 733
Иванисенко Т.В. 8, 733
Ивасенко Д.А. 5, 449
Игонина Т.Н. 5, 431
Игошин А.В. 3, 298
Илинский Ю.Ю. 6, 553
Ильичев А.А. 2, 214
Иолчиев Б.С. 4, 378

Кабиллов М.Р. 1, 86
Казанцева А.В. 2, 179
Казанцева А.Ю. 2, 159
Казачек А.В. 8, 798
Капустина И.С. 6, 568
Каргаполова К.Ю. 5, 422
Карлов А.В. 7, 704
Карпенко Л.И. 2, 214
Карссен Л.С. 4, 378
Касымова А.А. 5, 449
Керв Ю.А. 7, 630
Кириченко А.В. 4, 378
Кирьякова М.Н. 7, 609
Киселева Е.В. 2, 169
Киселева М.И. 6, 583
Клименко И.А. 4, 349
Ковалева О.Н. 6, 524
Коваль А.Д. 3, 240
Коваль О.А. 1, 14
Ковас Ю.В. 2, 179
Коломиец Т.М. 6, 583
Колосова И.В. 4, 394
Колчанов Н.А. 8, 719, 743
Комиссаров А.С. 8, 810
Конарев А.В. 7, 630
Конькова Н.Г. 7, 637
Корболина Е.Е. 1, 65
Королева М.Л. 6, 544

Костерин О.Э. 4, 359
Котикова А.И. 1, 50
Кочетов А.В. 2, 153; 3, 250
Кочиева Е.З. 6, 507
Кравцова Л.С. 1, 86
Кривина Е.С. 1, 74
Кручинина Ю.В. 7, 662
Кузнецова Л.И. 4, 385
Кузнецова М.В. 5, 486
Кузьмина Т.И. 3, 234
Кулакова А.В. 6, 507; 7, 652
Куликов И.М. 7, 622
Курочкин В.Е. 6, 544
Кучур П.Д. 8, 810

Лаврик И.Н. 8, 733
Ларичев К.Т. 3, 250
Ларкин Д.М. 3, 298
Лашина Н.М. 3, 272
Лебедкин Д.А. 8, 765, 773
Левинсон А.Л. 5, 431
Леонова И.Н. 7, 675
Лепехов С.Б. 2, 196
Липина Т.В. 4, 365
Лобакова Е.С. 6, 575
Лободина Е.В. 7, 645
Локачук М.Н. 4, 385
Лоскутов И.Г. 6, 524; 7, 597
Лукина К.А. 6, 524
Льбунь Е.В. 5, 477

Мавлютов Ю.М. 4, 349
Макарова Е.Н. 2, 159
Максютов Р.А. 4, 394
Малых С.Б. 2, 179
Мальшев Л.Л. 7, 630
Малькеева Д.А. 2, 169
Мандельштам М.Ю. 3, 319
Манучарова Н.А. 6, 575
Маркова Ю.А. 6, 568
Матвеева Т.В. 7, 697
Матора Л.Ю. 5, 422
Матушкин Ю.Г. 1, 96; 8, 719
Машкина Е.В. 1, 59
Мглинец А.В. 4, 359
Меркулова Т.И. 1, 65
Мешкова Л.В. 7, 609
Мещерякова Е.Н. 1, 109
Минушкина Л.С. 8, 758
Минчева Е.В. 1, 86
Миргазов Д.А. 6, 560
Мироненко Н.В. 3, 272
Митрофанова О.П. 2, 128
Михайловская В.С. 5, 486
Мищенко Е.Л. 2, 121
Моргунов А.И. 5, 413
Мотылева С.М. 7, 622
Муратова А.Ю. 5, 477
Мустафин Р.Н. 1, 40; 2, 179; 3, 308

- Назарова Г.Г. 4, 371
 Нгуен М.Л. 2, 146
 Немцев Б.Ф. 7, 662
 Немченко У.М. 5, 495
 Никифоров В.С. 1, 50
 Никонова А.А. 5, 467
 Новикова Л.Ю. 3, 264; 7, 599
 Нуждина Н.С. 1, 22
 Нурминский В.Н. 6, 568
 Нуштаева А.А. 1, 14
- Овчинникова Е.С. 7, 662
 Озерский П.В. 3, 264
 Озолина Н.В. 6, 568
 Окотруб С.В. 4, 365
 Омелянчук Н.А. 8, 721
 Осадчук А.В. 1, 96
 Осадчук Л.В. 1, 96
 Осипова Л.П. 2, 188
 Осянин К.А. 6, 560
 Ощепков Д.Ю. 1, 96
- Павловская Е.Н. 4, 385
 Пальянов А.Ю. 8, 780
 Панищева Д.В. 7, 622
 Панкратова Л.Ф. 6, 583
 Парахина О.И. 4, 385
 Пахотина И.В. 7, 609
 Пельтек С.Е. 1, 109
 Перетолчина Т.Е. 1, 86
 Пермязова Н.В. 4, 341
 Петрова Д.В. 4, 341
 Пискарев В.В. 7, 662
 Плеканчук В.С. 3, 281
 Подгаецкий М.А. 7, 622
 Подколотная О.А. 8, 798
 Подколотный Н.Л. 8, 798
 Полубоярова Т.В. 1, 109
 Пономаренко М.П. 1, 96; 3, 227; 8, 798
 Пономаренко П.М. 1, 96; 8, 798
 Попов В.С. 7, 630
 Поротников И.В. 2, 128
 Потоцкая И.В. 5, 413
 Притужалова А.О. 3, 234
 Прокудина О.И. 3, 281
 Проскурняк Л.П. 4, 371
 Пышная И.А. 1, 5
 Пышная О.Н. 7, 652
- Разгонова М.П. 7, 637
 Ракевич А.Л. 6, 568
 Рассказов Д.А. 8, 798
 Ревва П.М. 8, 733
 Ремезовская Н.Б. 5, 486
 Рихтер В.А. 1, 14
 Рожкова И.Н. 4, 365; 5, 431
 Ромашов Г.А. 3, 298
 Рудометов А.П. 2, 214; 3, 240
 Рудометова Н.Б. 2, 214
 Рудыч П.Д. 8, 773
- Рухлова Е.А. 3, 240
 Рябинин А.С. 6, 553
 Рязанова М.А. 3, 281
- Савилов Е.Д. 5, 495
 Савинкова Л.К. 3, 227; 8, 798
 Савкина О.А. 4, 385
 Савостьянов А.Н. 8, 765, 773
 Салина Е.А. 3, 250
 Сапрыгин А.Е. 8, 765, 773
 Свистунова Н.Ю. 7, 622
 Семенова Е.В. 7, 599
 Сергеева Е.М. 3, 250
 Серов О.Л. 4, 402
 Сибикеев С.Н. 6, 537
 Сидоренко А.Д. 8, 721
 Сидоров А.В. 6, 568
 Ситникова К.О. 5, 495
 Смирнов Д.Н. 1, 109
 Смоленская А.Е. 7, 630
 Смоленская С.Э. 7, 662
 Соколова Д.В. 1, 30
 Старич Эрьявец М. 5, 486
 Стоянов Х. 2, 139
 Супрун И.И. 7, 645
 Суслов В.В. 1, 96
 Сухарева Е.С. 5, 495
 Сухорева М.В. 5, 495
- Табиханова Л.Э. 2, 188
 Таможников С.С. 8, 773
 Татарская Д.Н. 3, 234
 Тахирова З.Р. 2, 179
 Темралеева А.Д. 1, 74
 Титов И.И. 8, 819, 826
 Тихонова Н.Г. 7, 630
 Тихонова О.А. 7, 630
 Ткаченко А.Г. 5, 458
 Ткаченко О.В. 5, 422
 Токмаков С.В. 7, 645
 Трегубчик Т.В. 4, 394
 Тумаева Т.А. 7, 622
 Тупикин А.Е. 1, 86
 Турковская О.В. 5, 477
 Турнаев И.И. 8, 787
- Уварова Е.А. 3, 327
 Ульданова Е.Е. 4, 365
 Ульянов А.В. 7, 704
- Фахрутдинов Н.А. 6, 560
 Федотов А.П. 1, 86
 Фёдорова С.А. 2, 169
 Филипенко М.Л. 2, 188
 Фирсов А.Б. 8, 826
 Фишман В.С. 8, 806
 Франк Ю.А. 5, 449
- Хаммадов Н.И. 6, 560
 Хандаев Б.М. 8, 798

Хатефов Э.Б. 7, 704
Хафизова Г.В. 7, 697
Хлебодарова Т.М. 8, 733, 743
Хлесткин В.К. 4, 385
Хлесткина Е.К. 7, 597
Храмеева Е.Е. 1, 109
Хуен Т.Н.Б.Т. 2, 146
Хуснутдинова Э.К. 1, 40; 2, 179
Хютти А.В. 3, 272

Цепилов Я.А. 4, 378
Цыганов И.В. 5, 458

Чадаева И.В. 1, 96; 3, 227; 8, 798
Челебиева Э.С. 3, 290
Чепурнов Г.Ю. 7, 662
Черняева Е.Н. 3, 298
Чешкова А.Ф. 2, 202
Чжоу Чэньси 7, 662
Чинь Д.М. 2, 146
Чистякова И.В. 3, 234
Чуйко Э.А. 4, 365
Чуркина Т.В. 2, 188

Шагимарданова Е.И. 1, 109
Шаманин В.П. 5, 413
Шамустакимова А.О. 4, 349

Шапиро Т.Н. 6, 575
Шарыпова Е.Б. 1, 96; 3, 227; 8, 798
Шашкова Т.И. 4, 378
Швалов А.Н. 4, 394
Шварцев А.А. 6, 544
Шеленга Т.В. 7, 630
Шереметьева М.Е. 8, 743
Шеховцов С.В. 1, 109
Шипова А.А. 1, 109
Широков А.А. 5, 422
Шихат О.В. 3, 290
Шишкина О.Д. 6, 553

Щелкунов С.Н. 4, 394
Щенникова А.В. 6, 507; 7, 652
Щербаков Д.Н. 2, 214; 3, 240
Щербаков Д.Ю. 1, 86
Щербань А.Б. 7, 684
Щёголев С.Ю. 5, 422

Юдин Н.С. 3, 298
Юсов В.С. 7, 609

Яковлева Т.В. 2, 159
Якубицкий С.Н. 4, 394
Яненко А.С. 8, 743

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”.

Регистрационное свидетельство ПИ № ФС77-45870 выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 20 июля 2011 г.

«Вавиловский журнал генетики и селекции» включен 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, Российский индекс научного цитирования, ВИНТИ, базы данных Emerging Sources Citation Index (Web of Science), Zoological Record (Web of Science), Scopus, PubMed Central, Ebsco, DOAJ, Ulrich's Periodicals Directory, Google Scholar, Russian Science Citation Index на платформе Web of Science, каталог научных ресурсов открытого доступа ROAD.

Открытый доступ к полным текстам:

на сайте ИЦиГ СО РАН – <https://sites.icgbio.ru/vogis/>

платформе Elpub – vavilov.elpub.ru/index.php/jour

платформе Научной электронной библиотеки – elibrary.ru/title_about.asp?id=32440

PubMed Central, <https://www.ncbi.nlm.nih.gov/pmc/journals/3805/>

Подписку на «Вавиловский журнал генетики и селекции» можно оформить в любом почтовом отделении России. Индекс издания 42153 по каталогу «Пресса России».

При перепечатке материалов ссылка на журнал обязательна.

✉ e-mail: 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.

Подписано в печать 20.12.2022. Выход в свет 30.12.2022. Формат 60 × 84 1/8. Усл. печ. л. 13.95.

Уч.-изд. л. 15.5. Тираж 150 экз. (1-й завод 1–45 экз.) Заказ № 360. Цена свободная.

Отпечатано в Сибирском отделении РАН, Морской проспект, 2, Новосибирск, 630090.