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

N.A. Kolchanov

Yu.G. Matushkin

Dear colleagues, dear readers!

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Science editors of this issue:

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

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