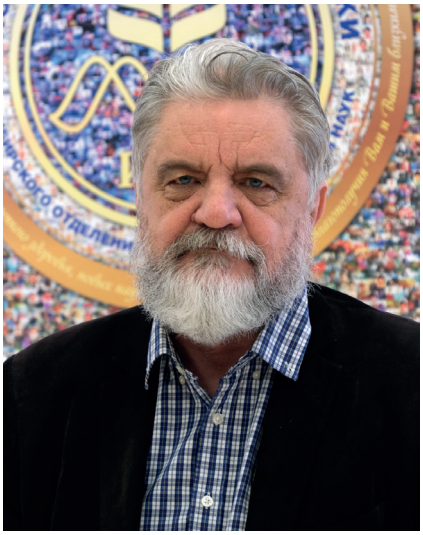
Вавиловский журнал генетики и селекции. 2023;27(7):725-727
DOI 10.18699/VJGB-23-84

Н.А. Колчанов



Ю.Г. Матушкин

Уважаемые коллеги, дорогие читатели!

Представляем вашему вниманию очередной выпуск «Вавиловского журнала генетики и селекции», посвященный биоинформатике и системной компьютерной биологии. Эти научные направления находятся сейчас в состоянии стремительной трансформации, что обусловлено вступлением наук о жизни в эпоху больших данных. Стремительное развитие омиксных технологий: геномики, транскриптомики, протеомики, метаболомики, а также других высокопроизводительных технологий изучения молекулярно-генетических основ функционирования живых систем привело к информационному взрыву в генетике, которая является основным источником больших данных в мировой науке, обогнав другие науки и технологии по скорости и объемам накопления экспериментальной информации.

Важнейшим результатом анализа, интерпретации, осмысливания больших генетических данных стало формирование новой парадигмы, в рамках которой главными объектами генетики являются не отдельные гены, а генные сети – группы координированно функционирующих генов, взаимодействующих друг с другом через свои продукты, такие как РНК, белки, метаболиты и другие вещества. Именно генные сети обеспечивают формирование всех фенотипических признаков организмов (молекулярных, биохимических, клеточных, физиологических, морфологических, поведенческих, ментальных и т. д.) на основе информации, закодированной в их геномах (Колчанов и др., 2000, 2013; Ananko et al., 2002).

Реконструкция генных сетей – очень сложная задача, требующая поиска, извлечения и интеграции информации, рассеянной среди десятков миллионов научных статей, тысяч фактографических баз данных и миллионов патентов, содержащих биологические, медицинские, фармакологические, химические и другие знания. Решение этой задачи потребовало разработки компьютерных программных систем для автоматизированного извлечения генетических знаний из упомянутых источников, использующих комбинацию традиционного анализа текста и методов машинного обучения (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2022). На сегодняшний день более 70 000 генных сетей и их основных компонентов (путей передачи сигналов, сетей белок-белковых, ДНК-белковых, РНК-белковых взаимодействий, метаболических путей) были реконструированы и представлены в базах данных (Pico et al., 2008; Caspi et al., 2020; Kanehisa et al., 2023).

Накопление больших данных привело к пониманию огромной сложности регуляции генных сетей на базовых уровнях их организации, проявляющейся в том, что любой элементарный фундаментальный биохимический или молекулярно-биологический процесс в генной сети контролируется, как правило, десятками, а иногда и сотнями элементарных регуляторных процессов, относится ли это к ферментативной активности белков, регуляции транскрипции генов или к «регуляции сложных метаболических путей» (Колчанов и др., 2008). Указанное обстоятельство создает огромные сложности при реконструкции молекулярных механизмов повреждающего

влияния геномной изменчивости на фенотипические характеристики организмов и клинические симптомы заболеваний, в том числе потому, что регуляторные процессы часто характеризуются высокой степенью нелинейности (Costanzo et al., 2019; Trifonova et al., 2021; Pratap et al., 2022) и динамической неустойчивостью по отношению к изменению начальных данных и констант физико-химических и молекулярно-биологических процессов, лежащих в основе функционирования генных сетей и регуляторных систем (Khlebodarova et al., 2018).

Обработка, анализ и интерпретация потоков больших генетических данных требуют разработки современных методов искусственного интеллекта, ориентированных на живые системы. Одним из ключевых событий, инициировавших в последние годы бурное развитие методов искусственного интеллекта, стала разработка новой архитектуры нейронных сетей, называемых трансформерами, ориентированных на обработку символьных последовательностей, включая тексты на естественных языках (Vaswani et al., 2017). Главная особенность трансформеров состоит в том, что порядок входных последовательностей при обработке не играет никакой роли. Это обеспечивает широкие возможности для распараллеливания, позволяя производить глубокое обучение моделей сразу на терабайтах данных, за гораздо меньшее время, чем было возможно ранее при классической архитектуре нейронных сетей.

Отметим несколько выдающихся достижений данного подхода. Важнейшее значение имеет создание качественных систем машинного перевода с одного естественного языка на другой (Jiao et al., 2023; Wang et al., 2023). Значение этого результата для науки, технологий, культуры, искусства, развития человеческих коммуникаций трудно переоценить.

На основе трансформерных моделей достигнут огромный успех в решении одной из центральных задач молекулярной биологии, над которой бились физики, химики, биологи в течение 60 лет, а именно в предсказании пространственной структуры глобулярных белков по их аминокислотным последовательностям. Для решения этой задачи были разработаны нейронные сети AlphaFold (Thornton et al., 2021) и Rosetta (https://www.rosettacommons.org/), предсказывающие 3D координаты тяжелых атомов белков с точностью, близкой к экспериментальной. Сеть была обучена на сотнях тысяч белков с известной пространственной структурой и десятках миллионов аминокислотных последовательностей.

Благодаря методам машинного обучения, использующим трансформерные подходы, открылась возможность моделирования динамики сложных молекулярно-биологических структур, содержащих очень большое (до 10^9) количество атомов (Pandey et al., 2022). Эти результаты имеют исключительное значение не только для фундаментальной науки, но и для широкого круга областей с громадным потенциалом фактического применения, таких как биотехнологии, генетика, медицина, фармакология, создание новых материалов и множество других.

После 2017 г., когда появились первые публикации по трансформерным технологиям, отмечена экспоненциальная динамика роста количества публикаций с использованием методов искусственного интеллекта (Eraslan et al., 2019; Boudry et al., 2022). Еще один подход к машинному обучению, получивший широкое распространение и развитие в последние годы, – это графовые нейронные сети (GNN), которые на основе векторного представления вершин графов с учетом их локального окружения дают качественно новые возможности для анализа сложных сетевых структур (Hamilton et al., 2017). Применение GNN эффективно для описания, анализа и моделирования широчайшего круга сетевых систем – как природных, так и антропогенных и технических: генных сетей, сетей межмолекулярных взаимодействий, сетей знаний, социальных и др. (Ektefaie et al., 2023).

В заключение следует отметить, что принципиальным ограничением для широкого применения методов искусственного интеллекта в практически значимых областях человеческой деятельности является непрозрачность принимаемых им решений. В ряде работ (Ma et al., 2018) показано, что стратегический путь преодоления этого недостатка – разработка гибридных информационных систем нового поколения, интегрирующих классические методы биоинформатики, системной компьютерной биологии и новые технологии искусственного интеллекта на основе онтологического описания предметных областей исследований. Только такой подход, как нам представляется, может обеспечить как скорость и качество обработки больших генетических данных с помощью методов искусственного интеллекта, так и прозрачность получаемых на его основе результатов.

Список литературы / References

Колчанов Н.А., Ананько Е.А., Колпаков Ф.А., Подколодная О.А., Игнатьева Е.В., Горячковская Т.Н., Степаненко И.Л. Генные сети. *Мол. биология*. 2000;34(4):533-544

[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)]

Колчанов Н.А., Гончаров С.С., Лихошвай В.А., Иванисенко В.А. Системная компьютерная биология. Новосибирск: Изд-во СО РАН, 2008

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

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;4(2):833-850

[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)]

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

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

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