

Реконструкция и статистические оценки генных сетей комплексных заболеваний на основе онлайн-инструментов биоинформатики

Волков И.А.¹, Анашкина А.А.^{1, 2}, Туркина В.А.¹, Савина Е.А.^{1, 2*}, Орлов Ю.Л.^{1, 3*}

¹ Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет), Москва, Россия

² Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва, Россия

³ Институт цитологии и генетики СО РАН, Новосибирск, Россия

* kaparos@list.ru, orlov@bionet.nsc.ru

Ключевые слова: медицинская информатика; медицинская генетика; онкология; комплексные заболевания; генные онтологии; генные сети; образование

Мотивация и цель: Исследование генных сетей комплексных заболеваний является важной задачей биомедицины, требующей интеграции данных. Бурный рост данных технологий секвенирования следующего поколения (NGS, next generation sequencing) требует разработки новых методов анализа, процессинга и предварительной обработки данных секвенирования с использованием математических и статистических методов, включая реализацию программ для клиник, для врачей, с удобным интерфейсом. Несмотря на существующий широкий набор компьютерных программ в мире, для решения конкретных задач анализа клинических данных в России необходима их адаптация, в том числе создание биоинформационных методик, учебных пособий и обучающих элементов для студентов, интернов и ординаторов медицинских учреждений [1, 2]. Отработка методики реконструкции генных сетей важна как для поиска генов-мишеней для терапии, так и для развития биоинформационных методов, поиска новых учебных решений для образования. Представлен научный проект исследования комплексных заболеваний, в которых сложно выделить генетические компоненты, включая в том числе ментальные расстройства, шизофрению, болезнь Паркинсона [3]. Практические цели проекта – разработка специализированных компьютерных программных модулей для реконструкции и анализа генных сетей комплексных заболеваний, которые могут быть использованы для задач цифрового образования и повышения квалификации, подготовки студентов биомедицинских и биоинженерных специальностей.

Методы и алгоритмы: Собраны имеющиеся программные инструменты, подготовлен конвейер обработки данных по созданию списков генов. Составление списков генов выполнялось на основе запросов к базам данных OMIM (Online Mendelian Inheritance in Man), GeneCards, MalaCards. Список уточнялся с использованием данных о дифференциальной экспрессии генов, включая не кодирующие РНК (из базы данных TCGA), и публикаций. Использовалась база данных геномов рака GDC Data Portal. Расчет генных онтологий производился с помощью открытых ресурсов для биоинформационного анализа: PANTHER (<http://pantherdb.org/>) и DAVID (<https://davidbioinformatics.nih.gov/>), для визуализации генных онтологий использовался ресурс g:GOSt (<http://biit.cs.ut.ee/gprofiler/gost>). Анализ ключевых

генов проводился с помощью базы данных GeneCards.org и каталога соматических мутаций при раке COSMIC (<https://cancer.sanger.ac.uk/cosmic/>).

Компьютерное исследование данных секвенирования выполняется на основе интеграции доступных данных секвенирования (RNA-seq и ChIP-seq из GEO NCBI) и компьютерных ресурсов: ArrayExpress, TCGA, ENCODE (ENCyclopedia Of DNA Elements), UCSC Genome Browser, а также российских компьютерных ресурсов TRRD, GeneNet, NPRD, SitEx и комплексов ICGenomics, ANDSystem. Использование данных секвенирования, представленных в GEO NCBI, выполнялось с помощью ресурса GEO Dataset Browser.

Результаты: В качестве основных примеров приложений рассмотрены задачи анализа опухолей мозга – для глиомы, менингиомы, с исследованием осложнений и сопутствующих заболеваний, связанных с инфекциями, в том числе по данным после пандемии коронавирусной инфекции, которые накоплены за последние годы. В качестве дополнительных научных приложений с использованием той же методики реконструкции функциональных сетей генов, белков и метаболитов выполнена компьютерная реконструкция генных сетей для ряда комплексных заболеваний: онкологических заболеваний – глиомы, рака молочной железы, колоректального рака, и ряда ментальных расстройств, таких как болезнь Паркинсона и ментальных расстройств [4]. При этом отработана методика анализа сети, оценки ее структуры, устойчивости к воздействию лекарственных соединений на гены-мишени, будут выполнены статистические оценки структуры генных сетей, оценки распределения числа связей, участия некодирующих РНК [5].

Представлен онлайн-ресурс для интеграции данных по ассоциированным генам комплексных заболеваний.

Выводы: Представлена методика биоинформационного исследования генов глиомы. Собран список ключевых генов заболевания, рассчитаны категории генных онтологий, построена сеть белок-белковых взаимодействий, выделены наиболее крупные кластеры генов. Рассмотрены структуры белков и потенциал белок-белковых взаимодействий для узловых генов рассмотренных генных сетей, таких как TP53, PTEN.

Финансирование: Исследование поддержано РНФ (грант 24-24-00563, «Разработка цифровых образовательных программ в биомедицине»).

Reconstruction and statistical estimates of gene networks of complex diseases based on online bioinformatics tools

Volkov I.A.¹, Anashkina A.A.^{1, 2}, Turkina V.A.¹, Savina E.A.^{1, 2*}, Orlov Y.L.^{1, 3*}

¹ Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia

² Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

³ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* kaparos@list.ru, orlov@bionet.nsc.ru

Key words: medical informatics; medical genetics; oncology; complex diseases; gene ontologies; gene networks; education

Motivation and Aim: The study of gene networks of complex diseases is an important biomedical task that requires data integration. The rapid growth of these next generation

sequencing technologies (NGS) requires the development of new methods for analyzing, processing and preprocessing sequencing data using mathematical and statistical methods, including the implementation of programs for clinics, for doctors, with a user-friendly interface. Despite the existing wide range of computer programs in the world, in order to solve specific tasks of analyzing clinical data in Russia, their adaptation is necessary, including the creation of bioinformatic techniques, textbooks and training elements for students, interns and residents of medical institutions [1, 2]. The development of methods for the reconstruction of gene networks is important both for the search for target genes for therapy, and the development of bioinformatic methods, the search for new educational solutions for the education of students based on available online bioinformatics tools. A scientific project for the study of complex diseases in which it is difficult to identify genetic components, including mental disorders, schizophrenia, and Parkinson's disease, is presented [3]. The practical goals of the project are the development of specialized computer software modules for the reconstruction and analysis of gene networks of complex diseases, which can be used for digital education and advanced training, training students of biomedical and bioengineering specialties, digital transformation of education.

Methods and Algorithms: The available software tools have been collected, a data processing pipeline for creating gene lists has been prepared. The list of genes was compiled based on current queries to the OMIM (Online Mendelian Inheritance in Man), GeneCards, and MalaCards databases. The list was refined using data on the differential expression of genes, including non-coding RNAs (from the TCGA database), and publications. The GDC Cancer genome database Data Portal was used. The calculation of gene ontologies was performed using open resources for bioinformatic analysis: PANTHER (<http://pantherdb.org/>) and DAVID (<https://davidbioinformatics.nih.gov/>), the g:GOS resource was used to visualize gene ontologies (<http://biit.cs.ut.ee/gprofiler/gost>). The analysis of key genes was carried out using a database GeneCards.org and the catalog of somatic mutations in cancer COSMIC (<https://cancer.sanger.ac.uk/cosmic/>).

The computer study of sequencing data is performed based on the integration of available sequencing data (RNA-seq and ChIP-seq from GEO NCBI) and computer resources: ArrayExpress, TCGA, ENCODE (ENCyclopedia Of DNA Elements), UCSC Genome Browser, as well as Russian computer resources TRRD, GeneNet, NPRD, SitEx and ICGenomics, ANDSystem complexes. The sequencing data presented in GEO NCBI was used using the GEO Dataset Browser resource.

Results: As the main examples of applications, the tasks of analyzing brain tumors are considered – for glioma, meningioma, with the study of complications and concomitant diseases associated with infections, including according to data after the coronavirus pandemic, which have accumulated in recent years. As additional scientific applications, using the same technique of reconstruction of functional networks of genes, proteins and metabolites, computer reconstruction of gene networks for a number of complex diseases – oncological diseases – glioma, breast cancer, colorectal cancer, and a number of mental disorders such as Parkinson's disease and metal disorders was performed [4].

At the same time, a methodology for analyzing the network, evaluating its structure, resistance to the effects of medicinal compounds on target genes has been developed, statistical estimates of the structure of gene networks, estimates of the distribution of the number of connections, and the participation of non-coding RNAs will be performed [5]. An online resource for integrating data on associated genes of complex diseases is presented.

Conclusion: The technique of bioinformatics research of glioma genes is presented. A list of key genes of the disease has been compiled, categories of gene ontologies have been calculated, a network of protein-protein interactions has been built, and the largest clusters of genes have been identified. The structures of proteins and the potential of protein-protein interactions for the nodal genes of the considered gene networks, such as TP53, PTEN, are considered.

Funding: The study was supported by the Russian Science Foundation (grant 24-24-00563).

Список литературы/References

1. Turkina V.A., Orlova N.G., Orlov Y.L. Biophysics education section and computational training discussion at VII Congress of Russian Biophysicists. *Biophys Rev.* 2023;15:807-809. doi 10.1007/s12551-023-01147-5
2. Anashkina A.A., Rubin A.B., Gudimchuk N.B. et al. VII Congress of Russian Biophysicists – 2023, Krasnodar, Russia. *Biophys Rev.* 2023;15:801-805. doi 10.1007/s12551-023-01164-4
3. Орлов Ю.Л., Галиева А.Г., Орлова Н.Г., Иванова Е.Н., Мозылева Ю.А., Анашкина А.А. Реконструкция геной сети болезни Паркинсона для поиска генов-мишеней. *Биомедицинская химия.* 2021;67(3):222-230. doi 10.18097/pbmc20216703222
[Orlov Y.L., Galieva A.G., Orlova N.G. et al. Reconstruction of gene network associated with Parkinson disease for gene targets search. *Biomeditsinskaya Khimiia.* 2021;67(3):222-230. doi 10.18097/pbmc20216703222 (in Russian)]
4. Voropaeva E.N., Pospelova T.I., Orlov Y.L., Churkina M.I., Berezina O.V., Gurazheva A.A., Ageeva T.A., Seregina O.B., Maksimov V.N. The Methylation of the p53 Targets the Genes MIR-203, MIR-129-2, MIR-34A and MIR-34B/C in the Tumor Tissue of Diffuse Large B-Cell Lymphoma. *Genes.* 2022;13(8):1401. doi 10.3390/genes13081401
5. Dergilev A., Orlova N., Dobrovolskaya O., Orlov Y. Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data. *J Integr Bioinf.* 2022;19(1):20200036. doi 10.1515/jib-2020-0036