

A COMPUTATIONAL PIPELINE FOR *DE NOVO* RECOGNITION OF TRANSCRIPTION FACTOR BINDING SITES IN BACTERIAL GENOMES

A. Mukhin, D. Oschepkov, S. Lashin

Kurchatov Genomic Center Institute Cytology and Genetics SB RAS,
630090, Novosibirsk, Russia
Institute Cytology and Genetics SB RAS,
630090, Novosibirsk, Russia
Novosibirsk State University,
630090, Novosibirsk, Russia

DOI: 10.24412/2073-0667-2024-4-69-83

EDN: UGUBKF

The search for transcription factor binding sites (TFBSs) in bacterial genomes is one of the most important steps for their study and subsequent use in biotechnology and microbiology. The characteristic length of TFBS is 5–20 nucleotide pairs, and each transcription factor has the ability to bind to a set of sites similar in sequence. The concept of motif is used to describe the spectrum of sequences that have substantial (non-random) similarity. That is, a motif in molecular biology is a group (or a representative of a group, depending on the context) of relatively short sequences of nucleotides (or amino acids) that have sufficient similarity due to their performance of a single biological function, e. g., binding of a single transcription factor. The similarity of motifs is directly used by various bioinformatics approaches for their de novo detection in genomic sequence samples, and is possible only if there is sufficient enrichment of the tested sample with the corresponding sequence similarity. In cases where the bacterial genome is insufficiently annotated, such as when working with a newly sequenced genome, it is the de novo motif detection method that proves to be the most effective for finding TFBSs. In this paper, we propose a set of computational motif search pipelines that take as input the bacterial genome data and its primary annotation. The proposed pipelines using two different approaches (full-genome search, when de novo motifs are searched for in a set of promoters of a single genome, and phylogenetic footprinting, when motifs are searched for among a set of promoters of similar genes and/or operons) to search for motifs, provide the researcher with a comprehensive set of settings for obtaining the most complete annotation by sites of both the whole genome and more detailed annotation of the regulatory region of the selected gene. The presented pipelines were implemented using both the modern Nextflow platform and scripts in the Python programming language. Also, the following tools were used within the pipelines: BoBro as a method for searching de novo motifs in promoters of a single organism; MP3, which implements de novo motif searching by phylogenetic footprinting in a set of promoters, GOST to identify similar genes and/or operons between two genome assemblies, OperonMapper to determine the operon structure of the genome, and TomTom for annotation of de novo motifs. We have developed an indexed metadata database for known bacterial genomes using an embedded SQLite DBMS, which allows us to significantly accelerate data retrieval for further calculations.

The work was supported by a budget project No FWNR-2022-0020.

Key words: pipeline, motifs, TFBS, genomics, Nextflow, Python, SQLite, JBrowse2, bioinformatics.

References

1. Seemann T. Prokka: rapid prokaryotic genome annotation // *Bioinformatics*. 2014. V. 30. N. 14. P. 2068–2069.
2. Pachkov M., Balwierz P. J., Arnold P., Ozonov E., Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates // *Nucleic Acids Research*. 2012. 11. V. 41. N D1. P. D214–D220. [El. res.]: <https://academic.oup.com/nar/article-pdf/41/D1/D214/3645388/gks1145.pdf>.
3. Robison K., McGuire A. M., Church G. M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome¹ Edited by R. Ebright // *Journal of Molecular Biology*. 1998. V. 284. N 2. P. 241–254. [El. res.]: <https://www.sciencedirect.com/science/article/pii/S002228369892160X>.
4. Dudek C.-A., Jahn D. PRODORIC: state-of-the-art database of prokaryotic gene regulation // *Nucleic acids research*. 2022. V. 50. N. D1. P. D295–D302.
5. Liu B., Zhang H., Zhou C., Li G., Fennell A., Wang G., Kang Y., Liu Q., Ma Q. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes // *BMC genomics*. 2016. V. 17. P. 1–12.
6. Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L., Jones R. T. Embryonic ϵ and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints // *Journal of molecular biology*. 1988. V. 203. N. 2. P. 439–455.
7. Yang J., Chen X., McDermaid A., Ma Q. DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses // *Bioinformatics*. 2017. V. 33. N 16. P. 2586–2588.
8. Bailey T. L., Johnson J., Grant C. E., Noble W. S. The MEME Suite // *Nucleic Acids Research*. 2015. 05. V. 43. N. W1. P. W39–W49. [El. res.]: <https://academic.oup.com/nar/article-pdf/43/W1/W39/17435890/gkv416.pdf>.
9. Sayers E. W., Bolton E. E., Brister J. R., Canese K., Chan J., Comeau D., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B., Pruitt K., Sherry S. Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2021. 12. V. 50. N D1. P. D20–D26. [El. res.]: <https://academic.oup.com/nar/article-pdf/50/D1/D20/42058080/gkab1112.pdf>.
10. Mukhin A. M., Kazantsev F. V., Klimenko A. I., Lakhova T. N., Demenkov P. S., Lashin S. A. The Web Platform for Storing Biotechnologically Significant Properties of Bacterial Strains // *International Conference on Parallel Computing Technologies* / Springer. 2021. P. 445–450.
11. Taboada B., Estrada K., Ciria R., Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes // *Bioinformatics*. 2018. 06. V. 34. N. 23. P. 4118–4120. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/34/23/4118/48921148/bioinformatics_34_23_4118.pdf.
12. Ma Q., Liu B., Zhou C., Yin Y., Li G., Xu Y. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*. 2013. 07. V. 29. N 18. P. 2261–2268. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/29/18/2261/50782707/bioinformatics_29_18_2261.pdf.
13. Bailey T. L. STREME: accurate and versatile sequence motif discovery // *Bioinformatics*. 2021. 03. V. 37. N 18. P. 2834–2840. [El. res.]: <https://academic.oup.com/bioinformatics/article-pdf/37/18/2834/50579626/btab203.pdf>.

-
14. Di Tommaso P., Chatzou M., Floden E. W., Barja P. P., Palumbo E., Notredame C. Nextflow enables reproducible computational workflows // *Nature biotechnology*. 2017. V. 35. N. 4. P. 316–319.
 15. Li G., Ma Q., Mao X., Yin Y., Zhu X., and Xu Y. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes // *Nucleic acids research*. 2011. V. 39. N. 22. P. e150–e150.
 16. Li G., Liu B., Ma Q., Xu Y. A new framework for identifying cis-regulatory motifs in prokaryotes // *Nucleic acids research*. 2011. V. 39. N 7. P. e42–e42.
 17. Mao X., Ma Q., Zhou C., Chen X., Zhang H., Yang J., Mao F., Lai W., Xu Y. DOOR 2.0: presenting operons and their functions through dynamic and integrated views // *Nucleic acids research*. 2014. V. 42. N. D1. P. D654–D659.
 18. Peltek S., Bannikova S., Khlebodarova T. M., Uvarova Y., Mukhin A. M., Vasiliev G., Scheglov M., Shipova A., Vasilieva A., Oshchepkov D., Bryanskaya A., Popik V. The Transcriptomic Response of Cells of the Thermophilic Bacterium *Geobacillus icigianus* to Terahertz Irradiation // *International Journal of Molecular Sciences*. 2024. V. 25. N 22.
 19. Diesh C., Stevens G. J., Xie P., De Jesus Martinez T., Hershberg E. A., Leung A., Guo E., Dider S., Zhang J., Bridge C., et al. JBrowse 2: a modular genome browser with views of synteny and structural variation // *Genome biology*. 2023. V. 24. N 1. P. 74.
 20. Pratt H., Weng Z. LogoJS: a Javascript package for creating sequence logos and embedding them in web applications // *Bioinformatics*. 2020. 03. V. 36. N 11. P. 3573–3575. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/36/11/3573/50670952/bioinformatics_36_11_3573.pdf.

ВЫЧИСЛИТЕЛЬНЫЙ КОНВЕЙЕР ПО РАСПОЗНАВАНИЮ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В БАКТЕРИАЛЬНЫХ ГЕНОМАХ *DE NOVO*

А. М. Мухин, Д. Ю. Ощепков, С. А. Лашин

Курчатовский Геномный Центр Института цитологии и генетики Сибирского отделения Российской академии наук (КГЦ ИЦиГ СО РАН),
630090, Novosibirsk, Russia
ФИЦ Институт цитологии и генетики СО РАН,
630090, Novosibirsk, Russia
Новосибирский государственный университет,
630090, Новосибирск, Россия

УДК 575.112

DOI: 10.24412/2073-0667-2024-4-69-83

EDN: UGUBKF

Задача поиска сайтов связывания транскрипционных факторов (ССТФ) в бактериальных геномах является одним из важнейших этапов их изучения и последующего использования в задачах биотехнологии и микробиологии. Характерная длина ССТФ — 5–20 пар нуклеотидов, и каждый транскрипционный фактор обладает способностью связываться с набором сайтов, сходных по последовательности. Поэтому поиск таких коротких последовательностей, имеющих достаточное, т. е. не случайное, сходство — т. н. мотивов — лежит в основе аннотации бактериальных геномов сайтами связывания. В статье описаны набор вычислительных конвейеров по поиску мотивов, которые принимают на вход данные бактериального генома и его первичной аннотации. Предлагаемые конвейеры, использующие два разных подхода (полногеномный поиск и филогенетический футпринтинг) к поиску мотивов, предоставляют исследователю исчерпывающий набор настроек для получения на выходе максимально полной аннотации сайтами как всего генома, так и более детально — регуляторного района выбранного гена. Представленные конвейеры реализованы как с использованием современной платформы Nextflow, так и скриптами на языке программирования Python. Разработанная нами индексируемая база метаданных для известных бактериальных геномов с использованием встраиваемой СУБД SQLite позволяет существенно ускорить извлечение данных для дальнейших расчетов.

Ключевые слова: конвейеры, мотивы, ССТФ, геномика, Nextflow, Python, SQLite, JBrowse2, биоинформатика, филогенетический футпринтинг.

Введение. Технологии высокопроизводительного секвенирования в молекулярной генетике дали существенный толчок развитию современных биотехнологий, обеспечив возможность массовой сборки бактериальных геномов для их анализа, модификации и дальнейшего использования отобранных штаммов бактерий в биотехнологических задачах,

Данная работа была поддержана бюджетным проектом FWNR-2022-0020.

например, для промышленного синтеза ферментов, белков медицинского и сельскохозяйственного назначения, лекарственных, профилактических и диагностических средств, незаменимых аминокислот и проч. На данный момент успешно применяются решения по аннотации вновь секвенированных геномов генами [1], однако задача их быстрой и эффективной аннотации регуляторными элементами до конца не решена. Аннотация бактериальных геномов и их конкретных регуляторных геномных последовательностей сайтами связывания транскрипционных факторов (ССТФ) является актуальной биологической задачей, поскольку работа клетки и производство определенных ферментов критическим образом зависят от существующих регуляторных геномных связей, реализующихся парой «транскрипционный фактор — сайт связывания». Как правило, ССТФ — это короткие участки ДНК от 5 до 20 пар оснований, узнаваемых соответствующими факторами транскрипции. Точное подтверждение связывания каждого отдельного сайта для конкретного ТФ и определенного штамма микроорганизма возможно только с помощью трудоемких экспериментальных методик. Данные по таким подтвержденным экспериментально ССТФ содержатся в *базах данных сайтов связывания транскрипционных факторов*, к таковым относятся, например, SwissRegulon [2], DPinteract [3] и PRODORIC [4]. Существенное разнообразие последовательностей, с которыми способен связываться каждый транскрипционный фактор каждого вида микроорганизмов, а также неполнота баз данных, содержащих известные ССТФ, не позволяет распознавать их в геномных последовательностях прямым сравнением с шаблоном и требует использования специализированных программ, основанных на различных эвристических подходах, использующих в том числе сходство между известными сайтами каждого ТФ.

Для описания спектра последовательностей, обладающих существенным (неслучайным) сходством, применяют понятие *мотива*. То есть мотив в молекулярной биологии — это группа (или представитель группы, в зависимости от контекста) относительно коротких последовательностей нуклеотидов (или аминокислот), обладающих достаточным сходством вследствие выполнения ими одной биологической функции, например, связывания одного транскрипционного фактора. На рис. 1 показаны набор последовательностей, содержащих сайты связывания транскрипционного фактора Lgr, и построенные на их основе обобщенные описания соответствующих им мотивов в виде позиционно-вероятностной матрицы и т. н. веб-лого — графического представления уровня консервативности в каждой позиции последовательности, описанного с использованием веса позиции. Сходство мотивов непосредственно используется различными подходами биоинформатики для их выявления *de novo* в выборках геномных последовательностей и возможно лишь при наличии достаточного обогащения тестируемой выборки соответствующими мотивами. В случае, когда бактериальный геном аннотирован недостаточно, например, в случае работы со вновь секвенированным геномом, именно метод выявления мотивов *de novo* оказывается наиболее эффективным для поиска ССТФ.

Как правило, ССТФ располагаются в т. н. промоторах — участках последовательностей перед стартами транскрипции регулируемых ими генов или оперонов — групп генов у бактерий, регуляция экспрессии которых осуществляется одним промотором. Поэтому, для выявления мотивов *de novo* в качестве тестируемой выборки, обогащенной сайтами связывания соответствующего ТФ, и используются выборки промоторов. При этом есть два варианта составления таких выборок: (1) возможно взять набор всех/некоторых промоторов конкретного исследуемого организма, либо (2) составить выборку из промоторов генов, ортологичных гену исследуемого организма из близкородственных организмов, где

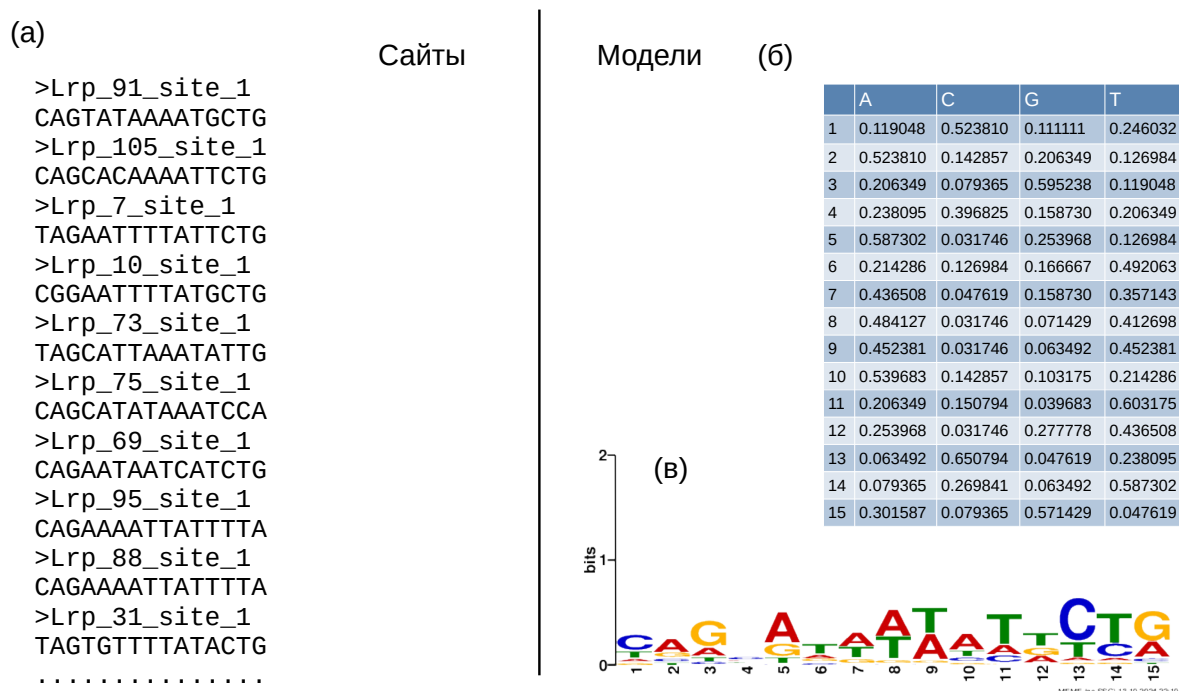


Рис. 1. Описание сайтов связывания транскрипционных факторов (ССТФ) и мотивов. На картинке (а) показано содержание файла с набором нуклеотидных последовательностей, ассоциированных с сайтом связывания (сокращенный набор), эти наборы можно обобщить в виде позиционно-вероятностной матрицы (б) и веб-лого (в), где по оси ординат отложены значения собственной информации для каждого нуклеотида в мотиве. В совокупности эти картинки описывают мотив Lrp

под ортологичными подразумеваются гены, которые у различных видов произошли от общего предшественника. Вариант поиска ССТФ с использованием выборок промоторов ортологичных генов близкородственных организмов носит название филогенетический фут-принтинг [5, 6].

Существующие подходы подразумевают достаточно обширную работу по составлению выборок для анализа, настройке и применению различных программ, систематизации и ранжированию полученных результатов. Однако существующие программные решения [7, 5, 8] содержат необходимые программные компоненты лишь частично, не позволяя комплексно и с минимальными трудозатратами выполнять массовый поиск ССТФ как во вновь секвенированных, так и в недостаточно изученных бактериальных геномах. Существенным ограничением в применимости перечисленных выше программных решений является отсутствие предоставляемого функционала по формированию необходимых входных данных — выборок промоторных областей генов-ортологов. Этот трудоемкий этап работ делегируется исследователю, то есть этапы поиска и отбора многочисленных генов-ортологов и извлечение их промоторных областей из соответствующих баз данных предполагается выполнять вручную. Аналогично, заключительный этап оценки схожести полученных *de novo* мотивов с наборами известных ССТФ из перечисленных выше БД требует предварительной подготовки результатов в соответствующем формате. Таким образом, была поставлена цель разработать автоматизированный конвейер для поиска сайтов связывания транскрипционных факторов в бактериальных геномах, включающий все

необходимые компоненты, в том числе, блоки для автоматического формирования двух вариантов выборок: как всех промоторов генома, так и промоторов генов-ортологов для проведения филогенетического футпринтинга, и обеспечивающий полный цикл анализа для конечного пользователя.

Результаты и Обсуждение. В данной работе реализован набор вычислительных конвейеров на разных платформах, позволяющий проводить полноценную аннотацию бактериальных геномов с помощью известных подходов *de novo* поиска ССТФ, выполняя следующие необходимые этапы анализа:

- 1) аннотирование генома оперонной структурой, необходимое для дальнейшего точного определения регуляторных/промоторных областей;
- 2) поиск *de novo* мотивов в выборке оперонов целевого генома;
- 3) функциональная аннотация вновь выявленных ССТФ.

Поиск *de novo* мотивов в промоторах целевого генома может осуществляться альтернативно с помощью двух подходов: либо в полной выборке промоторов целевого организма, либо на основании подхода филогенетического футпринтинга, осуществляя поиск ССТФ в наборе промоторов ортологичных (похожих) генов из одной таксономической группы с целевым организмом. В последнем случае необходимый список инструментов аннотации должен включать также:

— инструмент для поиска ортологичных генов заданного пользователем таксономического уровня;

— базы данных с последовательностями и аннотациями известных бактериальных геномов, что позволит автоматически осуществлять все необходимые операции по формированию требуемых выборок промоторов.

Такое комплексное решение подразумевает также наличие всех необходимых программных модулей, осуществляющих формирование требуемых для выполнения задачи выборок, операции по конвертации форматов, перенаправлению данных и последующему сохранению их части в соответствующие задаче хранилища. Такой комплексный подход позволит сократить до минимума затраты ресурсов на промежуточные, но требующие квалификации в программировании для персонала, проводящего биотехнологические исследования.

Для решения содержательных задач в области биоинформатики часто достаточно выполнить упорядочивание потока данных между существующими инструментами в программные конвейеры, и не требует разработки дополнительных новых алгоритмов. Несмотря на то, что число возможных конвейеров в задачах биоинформатики велико ввиду широкого спектра входных данных и поставленных научных задач, актуальность разработанных конвейеров определена известными задачами биотехнологии. Более того, разработанные конвейеры используют почти весь набор известных подходов для разметки бактериальных геномов сайтами связывания, и схема потока данных для этих подходов четко определена.

1. Конвейер подготовки входных данных. Два ниже описываемых и реализованных конвейера обладают общей частью по предобработке данных и получения оперонной структуры данных, оформленной в виде конвейера. В качестве входных данных конвейеры принимают пути до исследуемого генома в формате FASTA (рис. 1, а) и до файла с аннотацией генома в формате GFF. Последовательности в формате FASTA начинаются с однострочного описания, за которым следуют строки, содержащие собственно последовательность. Описание отмечается символом «>» в первой колонке. Файл в формате GFF —

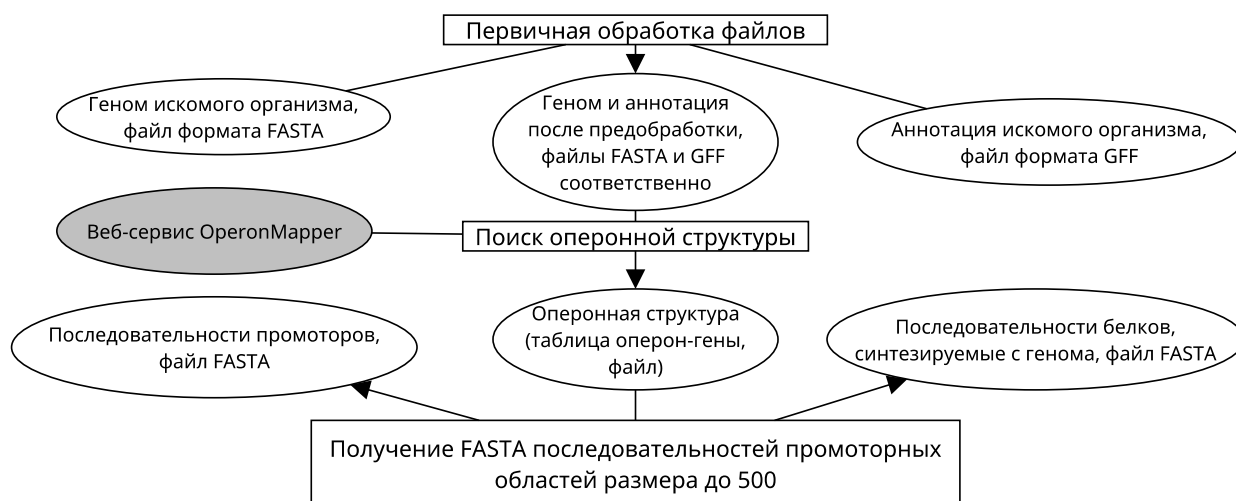


Рис. 2. Общая часть конвейера. Входными данными являются файлы генома искомого организма формата FASTA и аннотация генома формата GFF, выходными — файл с последовательностями промоторов формата FASTA и файл формата FASTA с последовательностями белков, синтезируемых с генома. Табличный файл ассоциации «оперон-гены» также может быть выходными данными

это текстовый файл, где для каждого функционального элемента генома отводится одна строка. Каждая строка содержит 9 полей (идентификатор хромосомы/последовательности нахождения, источник определения, тип элемента, начальные и конечные координаты, вес элемента, направление элемента относительно цепи, рамка считывания и атрибуты), разделенных знаком табуляции. После получения файлов проводится проверка этих данных на корректность и выполняется предобработка этих данных (обработка названий геномов, приведение значений в файле GFF в стандартный вид). Данный пункт необходим, так как в исследовании используются разные источники информации, это может быть база данных NCBI [9], или собственная база данных ЦГИМУ [10], или другие доступные БД, содержащие данные о бактериальных геномах.

Вторым этапом является определение оперонной структуры генома, то есть, какие гены из рассматриваемого генома объединены друг с другом под одним промотором, т. е. имеют общую регуляторную часть с общими ССТФ. Данный этап критически необходим для точного нахождения промоторов и определения их точных координат в геноме. Для выполнения этого этапа используется веб-сервис OperonMapper [11], с помощью которого можно проводить поиск оперонов для любых бактериальных геномов. В основе этого веб-сервиса лежит предобученная нейронная сеть, которая определяет оперонную структуру по похожим геномам из обучающей выборки. Для автоматизированной работы с ним была реализована программа на языке программирования Python с использованием библиотек requests и BeautifulSoup, которая отправляет HTTP-запрос к веб-сервису на выполнение анализа и далее, периодически, также проверяет статус задачи и в случае успеха получает ссылку на архив из HTML файла. После распаковки программой tar архива выходного файла, мы получаем текстовый табличный файл, где сначала записывается номер оперона, а далее — идентификаторы гена и их координаты, которые относятся к оперону. После получения данных по оперонной структуре, выполняется работа по выделению последовательностей промоторов и белковых последовательностей FASTA для каждого оперона.

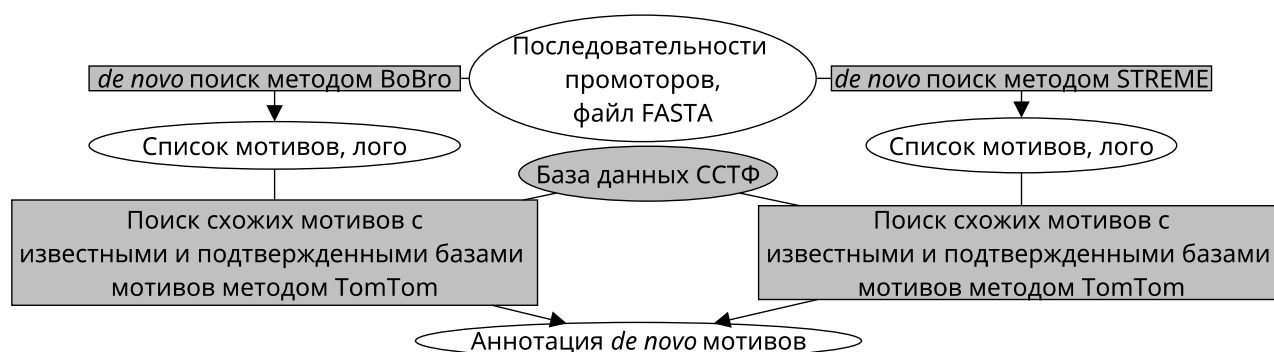


Рис. 3. Конвейер поиска мотивов *de novo* во всех промоторах. Серым цветом обозначены внешние программы и источники

Код этой части включен в оба ниже описываемых конвейера и может выполняться как независимым образом (т. к. представляет из себя набор скриптов на `bash` и `Python`) так и интегрироваться в существующие платформы для построения конвейеров, например, `bash` или `Nextflow`. На рис. 2 представлена графическая схема этого конвейера.

2. Вычислительный конвейер поиска мотивов *de novo* по всем промоторам генома. Первый конвейер для аннотации основан на подходе, осуществляющий поиск мотивов *de novo* во всех промоторных областях исследуемого генома.

Подход поиска *de novo* мотивов состоит в том, чтобы найти часто встречаемые мотивы в большом наборе промоторных областей или экспериментально определенных областей ДНК, где высока вероятность встретить сайт связывания соответствующего ТФ. Для анализа промоторов бактериальных геномов нами были использованы программы `BoBro` [12], `STREME` [13].

После этапа подготовки данных, конвейер, реализованный с использованием платформы `Nextflow` [14], осуществляет запуск выбранной программы поиска мотивов *de novo* и применяет программы для определения сходства между полученными мотивами с набором выборок известных ССТФ (рис. 3). Выборки известных бактериальных ССТФ содержатся в БД `SwissRegulon` [2], `DPinteract` [3] и `PRODORIC` [4], для сравнения найденных мотивов с этими выборками используется программа `Tomtom` из пакета `MEME` [8], для ее работы искомые БД сведены в текстовый файл с набором весовых матриц и их аннотаций. Файл с описанием конвейера для платформы `Nextflow` описан на специальном предметно-ориентированном языке (DSL), расширяющий язык программирования `Groovy`, и состоит из двух разделов: описание набора выполняемых процессов и описание потока данных, от обработки входных параметров и данных до конечной точки через описанные процессы. Описание процесса состоит из названия, описания входных и выходных данных либо через переменные, либо через маски файлов, и непосредственно кода выполняемого скрипта. Для каждой задачи анализ выполняется в специально созданной отдельной папке. Итоговый конвейер принимает на вход файлы с геномом в `FASTA` формате и аннотацией в `GFF` формате, а также строковый параметр «выбор программы поиска мотивов *de novo*».

Результаты, полученные этим конвейером, включают: потенциальные ССТФ и их координаты в исследуемом геноме, промоторные выборки целевого генома, используемые для сравнения с известными бактериальными ССТФ, список сходных известных бактери-

альных ССТФ. Именно эти данные крайне востребованы и могут быть использованы для решения различных задач биотехнологии и микробиологии, например, для оптимизации путей регуляции биосинтеза при создании штаммов-биопродукторов белков медицинского и сельскохозяйственного назначения, лекарственных, профилактических и диагностических средств, незаменимых аминокислот и проч.

3. Вычислительный конвейер поиска мотивов *de novo* с использованием методики филогенетического футпринтинга, база метаданных, лого последовательности. Методика филогенетического футпринтинга (англ. Phylogenetic footprinting) [5, 6] позволяет эффективно проводить поиск мотивов *de novo* для определенного оперона и использует последовательности промоторов для похожих или ортологичных генов (которые произошли от общего предка в процессе видообразования и выполняют схожие функции в организме) из других организмов. Промоторы в этом наборе должны быть достаточно эволюционно удаленными, чтобы стало возможным выявление сайтов связывания транскрипционных факторов, которые являются более эволюционно консервативными, на фоне менее консервативных участков промотора.

На входе конвейеру (рис. 4) подаются последовательность промоторов в исследуемом геноме, набор последовательностей белков, которые синтезируются с исследуемого генома, оперонная структура генома, параметр «номер исследуемого оперона» (число), а также необходимый уровень таксономии, требуемый для создания выборки промоторов, достаточно эволюционно удаленных для выявления ССТФ. Входные данные можно предварительно получить из выходных данных этапа предварительной обработки данных. Реализованный конвейер, выполняющий поиск мотивов *de novo*, состоит из следующих этапов (в табл. 1 описаны используемые программные инструменты и конвейеры):

- Отбор промоторов и последовательностей белков для целевого оперона;
- Получение последовательностей белков и оперонной структуры для геномов с определенным значением таксономии, взятые из базы метаданных SQLite и геномами из NCBI (об этом ниже);
- Поиск генов-ортологов между исследуемыми белками и белками, взятыми из базы метаданных SQLite. Этот этап выполняется с помощью программы GOST [15];
- Выделение последовательностей промоторов из генов-ортологов и формирование выборки последовательностей FASTA в определенном порядке (первый промотор — целевой, последующие — отобранные для генов-ортологов);
- Выполнение поиска мотивов *de novo* методом филогенетического футпринтинга. Данный этап выполняется с помощью программы MP3 [5];
- Поиск схожих мотивов с известными и подтвержденными базами ССТФ. Данный этап проводится с помощью программы TomTom и баз данных мотивов SwissRegulon, DPinteract и PRODORIC.

Конвейер поиска мотивов *de novo* методом филогенетического футпринтинга реализован в виде скрипта на языке программирования Python, который запускает последовательно вышеуказанные этапы конвейера, перенаправляя данные между программами и конвертируя форматы входных-выходных данных между промежуточными этапами.

Для быстрого и эффективного поиска данных по таксономии геномов, их оперонной структуре и по генам в соответствующих геномах с координатами была реализована база метаданных с использованием встраиваемой СУБД SQLite. Для отладки и пилотных запусков конвейера были отобраны и обработаны геномы (FASTA) для хорошо аннотированных бактериальных организмов с аннотациями (GFF) из базы данных NCBI [9] и оперон-

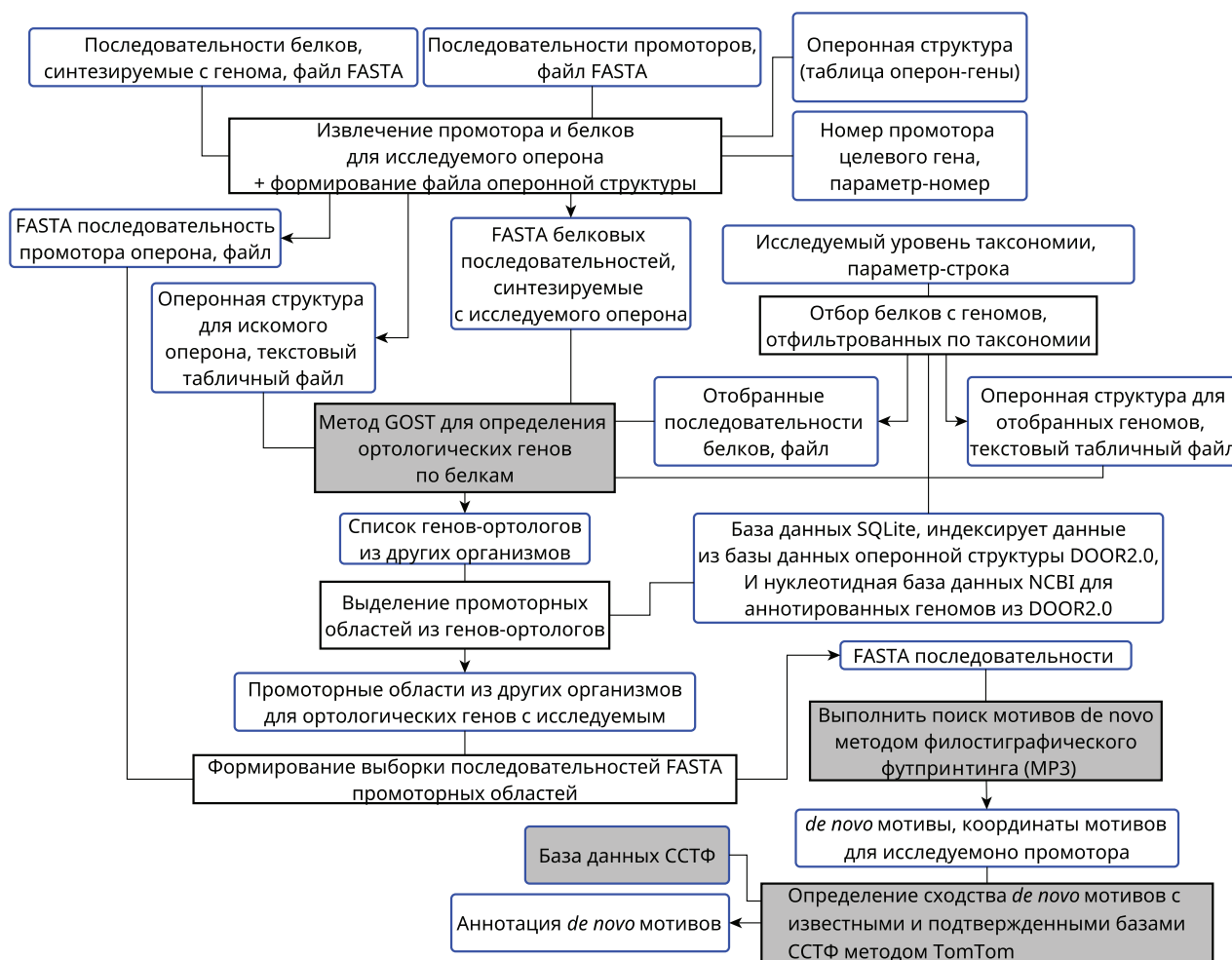


Рис. 4. Схема работы конвейера поиска мотивов *de novo* в промоторах генов-ортологов (метод филогенетического футпринтинга). Серым цветом обозначены внешние модули или данные, используемые в конвейере. Синий цвет границы элемента обозначает данные, серый цвет — выполняемый этап. Линия между двумя элементами означает «использование данных в этапе», стрелка — «этап производит следующие данные». Входными данными являются оперонная структура исследуемого гена, последовательности промоторов и белков, синтезируемые с исследуемого генома, а также параметры номер промотора для исследования (число) и требуемый для исследования уровень таксономии (строка)

ная разметка соответствующих геномов из базы данных DOOR2.0 [17]. Были реализованы скрипты на языках Bash и Python для загрузки данных из NCBI и DOOR2.0, обработки и сохранения этих данных в базу метаданных SQLite. Данные из DOOR2.0 представляют из себя JSON файлы для отображения на сайте и требовали постобработки, которая была реализована в виде скрипта на языке Python. Схема данной базы метаданных описана в приложении по ссылке: <https://disk.icgbio.ru/s/n8eJt55f7Q9oGDq>.

Таким образом, данный конвейер реализует метод филогенетического футпринтинга с выбором произвольного уровня филогенетического сходства, позволяет и более прецизионно осуществлять поиск потенциальных CCTF *de novo*, используя промоторные об-

Таблица 1

Описание используемых инструментов

Название	Описание	Ссылка
GOST	Получение пар ортологичных генов между двумя геномами	[15]
MP3	Конвейер для поиска мотивов с использованием подхода филогенетического футпринтинга	[5]
TomTom	Сравнение полученных <i>de novo</i> мотивов с базой известных и экспериментально подтвержденных мотивов	[8]
BoBro	Предсказание цис-регуляторных мотивов в наборе промоторов с помощью двойных выравниваний подстрок промоторных областях и методов над графами	[12, 16]
STREME	Поиск обогащенных мотивов в наборе последовательностей с помощью статистического теста Фишера	[13]
DOOR2.0	База данных предсказанных оперонов в бактериальных геномах	[17]
SwissRegulon	База данных полногеномной аннотации в регуляторных сайтах	[2]
DPinteract	База данных сайтов связывания для <i>E. coli</i>	[3]
PRODORIC	База данных геной регуляции	[4]

ласти генов-ортологов. Результаты, полученные этим конвейером, включают: потенциальные ССТФ и их координаты в исследуемом промоторе, их выборки из промоторов генов-ортологов, используемые для сравнения с известными бактериальными ССТФ, список сходных известных бактериальных ССТФ. Результаты работы этого конвейера также представляют значительную научную ценность, значительно расширяя спектр возможностей для анализа, предоставляемый первым конвейером поиска мотивов, описанным выше.

В качестве примера приведен результат работы описанного конвейера для поиска потенциальных ССТФ в бактериальном геноме *Geobacillus icigianus* (сборка NCBI_Assembly:GCA_000750005.2, загружена 28 марта 2024 г.) с помощью метода филогенетического футпринтинга (рис. 4), полученный на одном из этапов работы по анализу транскриптомного ответа *G. Icigianus* на терагерцовое излучение [18]. Показан участок промотора гена EP10_000119 со списком выявленных потенциальных ССТФ. Потенциальные ССТФ представлены в графическом виде в удобном для анализа биологами формате — т. н. лого последовательности (sequence logo). Он состоит из стопки букв алфавита (в случае ДНК — нуклеотидов) в каждой позиции. Относительные размеры букв указывают на их частоту в наборе последовательностей потенциальных ССТФ. Приведенное изображение (рис. 5) — результат отображения части генома в геномном браузере JBrowse2 [19], где лого генерируются в векторном формате SVG через реализацию плагина с использованием библиотек ReactJS и LogoJS [20]. Таким образом, реализация этого конвейера позволяет осуществлять распознавание ССТФ в отдельных промоторах исследуемого генома, используя концепцию использования больших геномных данных, благодаря чему позволяет повысить точность получаемых результатов и значительно расширяя спектр возможностей для анализа, предоставляемый первым конвейером поиска мотивов, описанным выше.

Заключение. Реализованные конвейеры и программные элементы позволяют снизить издержки научного сотрудника до нескольких часов без использования внешних ресурсов

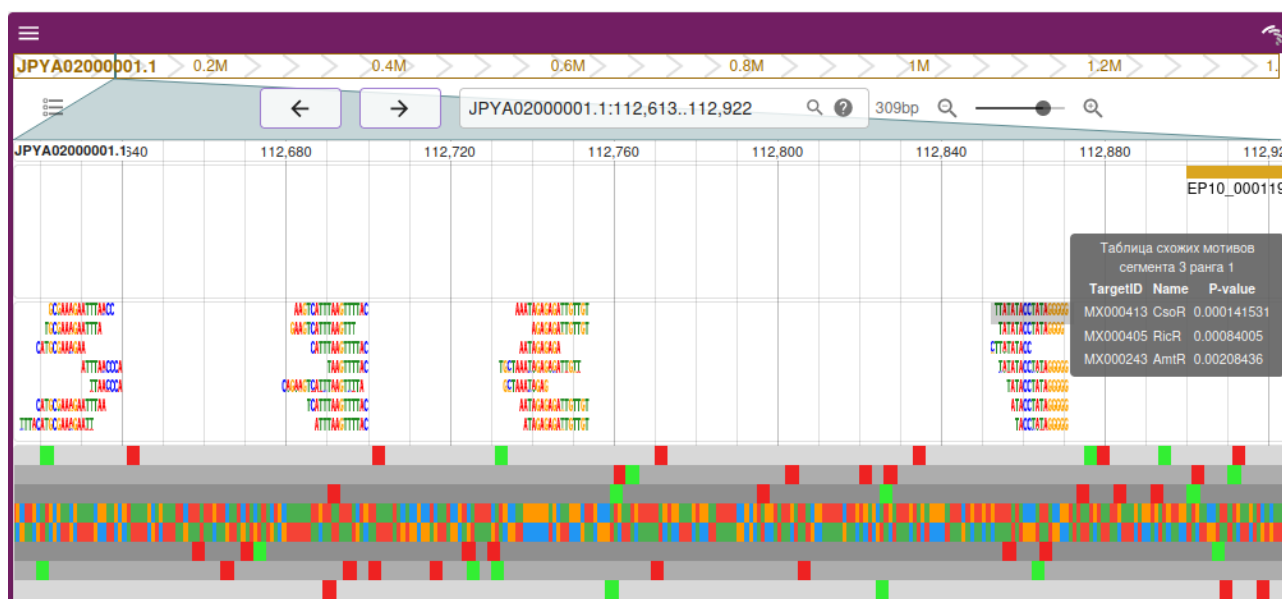


Рис. 5. Пример результата распознавания потенциальных ССТФ *de novo* с помощью конвейера филогенетического футпринтинга. Показана регуляторная область гена EP10_000119. Приведены выявленные с помощью программы MP3 мотивы, в выпадающем окне приведен список сходных известных бактериальных ССТФ с выделенным мотивом — результат работы программы TomTom. Отображение выполнено в программе JBrowse2

и с минимальными ручными действиями над данными со стороны пользователя. Также конвейеры позволяют автоматически насчитывать базу знаний по найденным мотивам и их аннотации в выбранных геномах, обеспечивая практически моментальный доступ исследователя к этим данным. Конвейеры можно запускать как в локальной среде, так и с использованием высокопроизводительного вычислительного кластера. Nextflow позволяет автоматически создавать список задач для системы управления заданиями Slurm высокопроизводительного кластера. Разработанная индексируемая база метаданных для известных бактериальных геномов с использованием встраиваемой библиотеки SQLite, интегрированная с программными компонентами, позволяет проводить быстрый поиск геномов по таксономии и генов по оперонам в выбранных геномах, что позволяет существенно ускорить поиск данных для дальнейших расчетов конвейера филогенетического футпринтинга. Отображение в геномном браузере JBrowse2 представляет интерактивный доступ к результатам работы конвейеров с интуитивно понятным отображением биологически значимой информации.

Список литературы

1. Seemann T. Prokka: rapid prokaryotic genome annotation // *Bioinformatics*. 2014. V. 30. N. 14. P. 2068–2069.
2. Pachkov M., Balwierz P. J., Arnold P., Ozonov E., Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates // *Nucleic Acids Research*. 2012. 11. V. 41. N D1. P. D214–D220. [El. res.]: <https://academic.oup.com/nar/article-pdf/41/D1/D214/3645388/gks1145.pdf>.

3. Robison K., McGuire A. M., Church G. M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. Edited by R. E. Eubright // *Journal of Molecular Biology*. 1998. V. 284. N 2. P. 241–254. [El. res.]: <https://www.sciencedirect.com/science/article/pii/S002228369892160X>.
4. Dudek C.-A., Jahn D. PRODORIC: state-of-the-art database of prokaryotic gene regulation // *Nucleic acids research*. 2022. V. 50. N. D1. P. D295–D302.
5. Liu B., Zhang H., Zhou C., Li G., Fennell A., Wang G., Kang Y., Liu Q., Ma Q. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes // *BMC genomics*. 2016. V. 17. P. 1–12.
6. Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L., Jones R. T. Embryonic ϵ and γ globin genes of a prosimian primate (*Galago crassicaudatus*): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints // *Journal of molecular biology*. 1988. V. 203. N. 2. P. 439–455.
7. Yang J., Chen X., McDermaid A., Ma Q. DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses // *Bioinformatics*. 2017. V. 33. N 16. P. 2586–2588.
8. Bailey T. L., Johnson J., Grant C. E., Noble W. S. The MEME Suite // *Nucleic Acids Research*. 2015. 05. V. 43. N. W1. P. W39–W49. [El. res.]: <https://academic.oup.com/nar/article-pdf/43/W1/W39/17435890/gkv416.pdf>.
9. Sayers E. W., Bolton E. E., Brister J. R., Canese K., Chan J., Comeau D., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B., Pruitt K., Sherry S. Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2021. 12. V. 50. N D1. P. D20–D26. [El. res.]: <https://academic.oup.com/nar/article-pdf/50/D1/D20/42058080/gkab1112.pdf>.
10. Mukhin A. M., Kazantsev F. V., Klimenko A. I., Lakhova T. N., Demenkov P. S., Lashin S. A. The Web Platform for Storing Biotechnologically Significant Properties of Bacterial Strains // *International Conference on Parallel Computing Technologies* / Springer. 2021. P. 445–450.
11. Taboada B., Estrada K., Ciria R., Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes // *Bioinformatics*. 2018. 06. V. 34. N. 23. P. 4118–4120. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/34/23/4118/48921148/bioinformatics_34_23_4118.pdf.
12. Ma Q., Liu B., Zhou C., Yin Y., Li G., Xu Y. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*. 2013. 07. V. 29. N 18. P. 2261–2268. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/29/18/2261/50782707/bioinformatics_29_18_2261.pdf.
13. Bailey T. L. STREME: accurate and versatile sequence motif discovery // *Bioinformatics*. 2021. 03. V. 37. N 18. P. 2834–2840. [El. res.]: <https://academic.oup.com/bioinformatics/article-pdf/37/18/2834/50579626/btab203.pdf>.
14. Di Tommaso P., Chatzou M., Floden E. W., Barja P. P., Palumbo E., Notredame C. Nextflow enables reproducible computational workflows // *Nature biotechnology*. 2017. V. 35. N. 4. P. 316–319.
15. Li G., Ma Q., Mao X., Yin Y., Zhu X., and Xu Y. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes // *Nucleic acids research*. 2011. V. 39. N. 22. P. e150–e150.
16. Li G., Liu B., Ma Q., Xu Y. A new framework for identifying cis-regulatory motifs in prokaryotes // *Nucleic acids research*. 2011. V. 39. N 7. P. e42–e42.

17. Mao X., Ma Q., Zhou C., Chen X., Zhang H., Yang J., Mao F., Lai W., Xu Y. DOOR 2.0: presenting operons and their functions through dynamic and integrated views // *Nucleic acids research*. 2014. V. 42. N. D1. P. D654–D659.

18. Peltek S., Bannikova S., Khlebodarova T. M., Uvarova Y., Mukhin A. M., Vasiliev G., Scheglov M., Shipova A., Vasilieva A., Oshchepkov D., Bryanskaya A., Popik V. The Transcriptomic Response of Cells of the Thermophilic Bacterium *Geobacillus icigianus* to Terahertz Irradiation // *International Journal of Molecular Sciences*. 2024. V. 25. N 22.

19. Diesh C., Stevens G. J., Xie P., De Jesus Martinez T., Hershberg E. A., Leung A., Guo E., Dider S., Zhang J., Bridge C., et al. JBrowse 2: a modular genome browser with views of synteny and structural variation // *Genome biology*. 2023. V. 24. N 1. P. 74.

20. Pratt H., Weng Z. LogoJS: a Javascript package for creating sequence logos and embedding them in web applications // *Bioinformatics*. 2020. 03. V. 36. N 11. P. 3573–3575. [El. res.]: https://academic.oup.com/bioinformatics/article-pdf/36/11/3573/50670952/bioinformatics_36_11_3573.pdf.



Мухин Алексей Максимович — младш. науч. сотрудник Института цитологии и генетики СО РАН, E-mail: mukhin@bionet.nsc.ru.

Мухин Алексей Максимович окончил ФИТ НГУ в 2019 году, в 2023 — аспирантуру ИЦиГ СО РАН. С 2017 г. сотрудник ИЦиГ СО РАН. В сфере его научных интересов — программное обеспечение в области биологии.

Mukhin Aleksey Maksimovich graduated from Faculty of Information Technology of the Novosibirsk State University in 2019, in 2023 — postgraduate program of ICG SB RAS. Since 2017 he has been an employee of ICG SB RAS. His research interests include software in the field of biology.



Ощепков Дмитрий Юрьевич — канд. биол. наук, старш. науч. сотрудник Института цитологии и генетики СО РАН, E-mail: diman@bionet.nsc.ru.

Ощепков Дмитрий Юрьевич окончил ФФ НГУ в 1999 году. С 1998 г. работает в ИЦиГ СО РАН. В сфере его научных интересов — компьютерная геномика и транскриптомика. Автор более 60 работ.

Dmitry Yurievich Oshchepkov graduated from the Faculty of Physics of the Novosibirsk State University in 1999. Since 1998, he has been an employee of the Institute of Cytology and Genetics of the Russian Academy of Sciences. His research interests include computational genomics and transcriptomics. He is the author of more than 60 scientific articles.



Лашин Сергей Александрович — канд. биол. наук, доцент, ведущ. науч. сотр. сектора биоинформатики и информационных технологий в генетике Института цитологии и генетики СО РАН, E-mail: lashin@bionet.nsc.ru.

Окончил в 2003 г. ММФ НГУ. Специалист в области математического и компьютерного моделирования биологических систем широкого круга — молекулярно-генетических, популяционно-генетических, экологических, разработки биоинформатических методов, программного обеспечения и баз данных.

Graduated in 2003 from MMF NSU. Specialist in the field of mathematical and computer modeling of biological systems of wide range — molecular-genetic, population-genetic, ecological, development of bioinformatic methods, software and databases.

Дата поступления — 07.06.2024