

СЕКЦИЯ «БИОИНФОРМАТИКА»

ВЕБ-ИНСТРУМЕНТ ДЛЯ ЛОКАЛЬНОГО ПОИСКА САЙТОВ СВЯЗЫВАНИЯ РЕГУЛЯТОРОВ ТРАНСКРИПЦИИ В ПОСЛЕДОВАТЕЛЬНОСТЯХ БАКТЕРИАЛЬНЫХ ГЕНОМОВ

А.В. Дигрис, П.В. Вычик, Е.И. Дувалов,
В.В. Скакун, Е.А. Николайчик

Белорусский государственный университет, Минск, Беларусь,
E-mail: digris@bsu.by

Представлен веб-инструмент, предназначенный для локального поиска регуляторных последовательностей в бактериальных геномах. Работа инструмента базируется на применении моделей регуляторных мотивов из базы данных BacRegDB, ключевой особенностью которых является использование CR-тегов, контролирующих корректность применения моделей к анализу конкретного генома. Новый инструмент существенно дополняет доступные на веб-портале BacRegDB программные средства анализа регуляторной информации, содержащейся в геномных последовательностях. Работа инструмента продемонстрирована на примере анализа генома *Pectobacterium versatile* 3-2.

Ключевые слова: регуляция транскрипции; регуляторные мотивы; BacRegDB; базы данных; веб-приложение.

Введение. Использование хранящихся в нуклеотидных базах данных бактериальных геномных последовательностей в практических целях, например, при конструировании штаммов-продуцентов или в фундаментальных исследованиях регуляторных процессов, серьезно затруднено отсутствием аннотации регуляторных элементов (промоторов, операторов, терминаторов и др.). Для решения этой проблемы мы разработали программное обеспечение для аннотации известных регуляторных элементов в бактериальных геномах [1], а позднее предложили подход к идентификации операторных последовательностей *de novo* в совершенно неисследованных геномах [2]. Этот подход использует концепцию ярлыков из критичных, т. е. непосредственно контактирующих с азотистыми основаниями ДНК, аминокислотных остатков транскрипционных факторов (т. н. CR-тегов). Идентичность CR-тегов двух транскрипционных факторов одного семейства означает, что они имеют одинаковую специфичность распознавания ДНК и один и тот же мотив сайта связывания. Поэтому при совпадении CR-тега транскрипционного фактора исследуемого организма с CR-тегом известного транскрипционного фактора применение математической модели операторного мотива известного транскрипционного фактора к исследуемому геному будет корректным.

Для структурированного и целостного хранения регуляторных мотивов, ассоциированных с CR-тегами, была разработана реляционная база данных (БД) [3], а для обеспечения удобного доступа пользователей к этой базе данных организован web-портал BacRegDB. В дополнение к организации графического интерфейса доступа к базе данных, BacRegDB предоставлял пользователям два аналитических инструмента: 1) идентификатор и классификатор транскрипционных факторов и 2) средство полногеномной аннотации операторных элементов [4]. Вторым инструментом наиболее полезен для практикующих биологов и предполагает последовательное применение операторных профилей по одному к одной и той же геномной последовательности для получения максимально полной картины транскрипционной регуляции исследуемого организма. Однако чаще встречается другая практическая задача: анализ транскрипционной регуляции одного гена или оперона. В этом случае трудоемкий и времязатратный полногеномный анализ является избыточным, а более целесообразным видится организация локального поиска потенциальных транскрипционных регуляторов гена-мишени. Разработка такого инструмента – локального сканера потенциальных регуляторных мотивов, – и явилась целью настоящей работы.

Веб-инструмент. Алгоритм работы веб-инструмента «локальный сканер» основан на процедуре поиска регуляторной информации с применением программ hmmscan и nhmmscan [5] и последующей обработке их результатов с помощью скрипта, написанного на языке Python. Выбор языка Python преимущественно определялся наличием удобного функционала обработки геномных данных библиотеки Biopython [6]. Запуск локального сканера осуществляется серверной частью веб-приложения, на основе которого работает портал BacRegDB, в ответ на запрос, пришедшей от клиента. Входными данными для инструмента являются файл генома в формате .gbk и диапазон координат, в котором будет выполняться поиск. Для обеспечения приемлемого времени работы локального сканера максимальный размер регуляторной области ограничен 2000 н.п.

Локальный сканер имеет два режима работы: строгий, когда требуется отбор только тех профилей операторных мотивов, для которых CR-тег профиля соответствует множеству CR-тегов обрабатываемого генома, и нестрогий; когда такой отбор не требуется. Режим работы инструмента определяется пользователем. На первом шаге алгоритма средствами библиотеки Biopython из .gbk файла по заданным координатам экстрагируется фрагмент генома. Данный фрагмент генома в виде файла в формате FASTA подается на вход программы nhmmscan, реализующей процедуру поиска операторов. При выполнении процедуры поиска применяется биб-

лиотека из всех скрытых марковских моделей операторных мотивов, имеющих в базе данных BacRegDB [3]. При работе в строгом режиме реализуется фильтрация результатов проведенного поиска, для чего по описанному в [2] принципу формируется множество CR-тегов транскрипционных факторов, соответствующих обрабатываемому геному. Результаты работы локального сканера сохраняются в формате JSON для последующей передачи в клиентскую часть веб-приложения с целью визуализации.

Локальный сканер реализован на веб-портале BacRegDB и доступен по адресу <http://bacregdb.bsu.by/tools?toolName=scan>. Для демонстрации работы веб-инструмента выполнен поиск локальных регуляторов в геноме *Pectobacterium versatile* 3-2 в диапазоне координат 1257200-1257899 н.п. Полученные результаты представлены на рисунке.

The screenshot shows the BacRegDB web application interface. At the top, there is a navigation bar with links: Home, TF Families, Tools, About, Contacts, Links. A search bar is also present. Below the navigation bar, the page title is 'Analysis tools'. A message says 'Select your .gbk file and apply tools bellow for the analysis.' Below this, there is a section for 'File name (.gbk):' with a 'Choose File' button and a file named 'Pve32_2023.09.gbk'. There are three tabs: 'Annotation', 'TF classifier', and 'TFBS scan'. The 'TFBS scan' tab is selected, showing 'Scan results for Pve32_2023.09.gbk'. The results are displayed in a table with columns: TF, Locus Tag, Operator sequence, Coordinates, Strand, E-value, and Bit score. The table contains six rows of results. Below the table, it says 'Total TF count: 6'. There are two buttons: 'Download as TSV file' and 'New scan'. At the bottom, there is a footer with 'Addresses:' (Belarusian State University, 4 Nezavisimosti Avenue, Minsk, 220030, Belarus.) and 'Cookie Policy:' (Our web site does not store or process cookies.). The footer also includes '©BacRegDB - 2024. Belarusian State University.'

TF ↑↓	Locus Tag ↑↓	Operator sequence ↑↓	Coordinates ↑↓	Strand ↑↓	E-value ↑↓	Bit score ↑↓
NRGTRK_Fis	OA04_02890	ATTGGTGATATTTTAACCAT	1257619:1257638	+	0.24	10.4
NRGTRK_Fis	OA04_02890	ATGGTTAAATATCACCAAT	1257638:1257619	-	0.24	10.4
PDRTDVTRR_ArcA	OA04_39400	ATTTTATCAAATTGATAAA	1257463:1257483	+	1.8	8.1
PDRTDVTRR_ArcA	OA04_39400	ATTTGATAAAAAATTATATAA	1257476:1257456	-	1.9	8.0
RSHTDVLGR_PhoP	OA04_23960	TATTATATTTTT	1257440:1257453	+	0.97	8.9
RSHTDVLGR_PhoP	OA04_23960	TTATTTAATAATTTTTAT	1257453:1257471	+	2.5	7.7

Результаты поиска регуляторных последовательностей для генома *Pectobacterium versatile* 3-2 в диапазоне координат 1257200-1257899 н.п.

В указанном интервале расположена регуляторная область дивергона *pell/pehA*, кодирующего два фактора вирулентности этой патогенной бактерии. Сканер успешно идентифицировал два сайта связывания глобального регулятора PhoP, чье присутствие в этой области было недавно продемонстри-

ровано экспериментально [7]. Также идентифицированы два сайта связывания регулятора ответа ArcA, ответственного за адаптацию к анаэробнозису, и один сайт для глобального регулятора Fis. Для Fis единственный сайт указан дважды, поскольку использован симметричный (палиндромный) профиль, а сканирование выполнено для обеих цепей ДНК.

Выводы. Разработанный веб-инструмент для локального поиска регуляторных последовательностей в бактериальных геномах, основанный на использовании CR-тегов, позволяет за короткое время определить потенциальные регуляторы конкретного бактериального гена или оперона, что очень полезно при конструировании штаммов-продуцентов и фундаментальных исследованиях транскрипционной регуляции отдельных генов/оперонов. Создание данного инструмента позволило расширить линейку средств анализа геномных данных доступных для широкого круга исследователей на веб-портале BacRegDB по адресу <http://bacregdb.bsu.by/tools>.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Nikolaichik Y., Damienikan A.U. Sigmoid: a user-friendly tool for improving bacterial genome annotation through analysis of transcription control signals. // PeerJ 4. 2016. P. e2056. DOI: <https://doi.org/10.7717/peerj.2056>
2. Nikolaichik Y., Vychik P. New approach to genome-wide automated inference of bacterial transcription factor binding sites // Abstracts of the Twelfth International Multiconference "Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020)" 06-10 July 2020, Novosibirsk: Institute of Cytology and Genetics. – P. 75-76. DOI: <https://doi.org/10.18699/BGRS/SB-2020-046>
3. Скакун В.В., Николайчик Е.А. Разработка базы данных мотивов регуляции транскрипции у бактерий. // Информатика. Т. 19. 2022. С. 59-71. DOI: <https://doi.org/10.37661/1816-0301-2022-19-1-59-71>
4. Дигрис А.В., Вычик П.В., Дувалов Е.И., Скакун В.В., Николайчик Е.А. BacRegDB – база данных и набор инструментов для работы с регуляторной информацией в бактериальных геномах. // Материалы 7-й Международной научно-практической конференции «Прикладные проблемы оптики, информатики, радиофизики и физики конденсированного состояния» 18-19 мая 2023 г., Минск / отв. ред. Кольчевская И.Н. 2023. С. 158-160.
5. Wheeler T.J., Eddy S.R. nhmmer: DNA homology search with profile HMMs. // Bioinformatics. Vol. 29. 2013. P. 2487-2489. DOI: <https://doi.org/10.1093/bioinformatics/btt403>
6. Biopython: freely available Python tools for computational molecular biology and bioinformatics. / Cock P.J. [et al.] // Bioinformatics. 2009. Vol. 25. P.1422-1423.
7. Kravchenko U., Gogoleva N., Kolubako A., Kruk A., Diubo J., Gogolev Y., Nikolaichik Y. The PhoPQ two-component system is the major regulator of cell surface properties, stress responses and plant-derived substrate utilisation during development of *Pectobacterium versatile*-host plant pathosystem. // Frontiers in Microbiology. 2021. Vol. 11. P. 3596. DOI: <https://doi.org/10.3389/fmicb.2020.621391>