УДК 577.21, 57.081.23

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

Белорусский государственный университет, г. Минск, Республика Беларусь

В работе предложены инструменты для анализа регуляторной информации в бактериальных геномах, выполняющие аннотирование сайтов связывания отдельных транскрипционных факторов, а также классификацию транскрипционных факторов на основе скрытых марковских моделей, в том числе и собственной разработки. Инструменты доступны для использования через веб-интерфейс новой базы данных регуляторной информации BackRegDB. Работоспособность и эффективность инструментов подтверждена на примере анализа геномных последовательностей бактерий разных таксономических групп.

Введение. На современном этапе развития геномных технологий актуальной задачей является обработка больших объемов информации, получаемой в ходе геномного секвенирования. При исследовании геномов бактерий важную роль играет их максимально полная функциональная аннотация, состоящая в идентификации кодирующих последовательностей, регуляторных элементов, рРНК и тРНК, транспозонов и других функциональных единиц генома. Регуляторные элементы при экспрессии генома живого организма позволяют ему реагировать на стимулы внешней среды и поддерживать клеточный гомеостаз. Основным механизмом такой адаптивности является дифференциальная экспрессия генов. Возможность активировать или репрессировать транскрипцию конкретных групп генов реализуется преимущественно путем взаимодействия с их регуляторными областями особого класса ДНК-связывающих белков – транскрипционных факторов (ТФ). Восстановление информации о регуляторных сетях в бактериальных геномах путем идентификации кодируемых ТФ, поиска их сайтов связывания и контролируемых транскрипционных единиц, определение условий, являющихся внешним сигналом для изменения экспрессии, – все это является основной предпосылкой для решения многих фундаментальных и прикладных задач биотехнологии, и делает важным создание инструментов для полноценной аннотации регуляторной информации. Существующие программы-аннотаторы не используют информацию о регуляторных мотивах и сайтах связывания, накопленную в общедоступных базах данных. Более того, большинство ресурсов не предлагает иных подходов к применению этой информации для новых геномов, кроме как принципа достаточного уровня гомологии аминокислотной последовательности транскрипционного регулятора. Данные факты делают особенно актуальной задачу по созданию инструмента для аннотации регуляторной информации, особенно в контексте более точного критерия применимости известных регуляторных мотивов в новых геномах.

В ходе нашей работы были систематизированы сведения об известных сайтах связывания и регуляторных мотивах, депонированных в таких базах данных как RegulonDB, CollecTF, RegPrecise, Prodoric2. В результате были получены новые регуляторные мотивы высокого качества на основе доступной экспериментально информации о сайтах связывания и результатов работы de-novo конвейера программы SigmoID 1 для поиска регуляторных мотивов. Для удобства использования регуляторных мотивов ТФ реализована база данных BacRegDB [2, 3] и разрабатывается веб-портал для доступа к хранящейся в ней информации. Инструменты веб-портала способны выполнять аннотирование сайтов связывания определенных ТФ, идентификацию и классификацию ТФ в геноме на основе скрытых марковских моделей. Ключевой особенностью регуляторных мотивов BacRegDB является использование тега из критических аминокислотных остатков (CR-тег) – последовательности аминокислот, располагающихся в позициях ДНК-связывающего домена, которые определяют специфичность узнаваемой последовательности ДНК. Определение CR-тега возможно для ТФ в загружаемых пользователем геномах, что определяет возможные для использования регуля-

торные мотивы в базе данных. Работоспособность инструментов подтверждена при анализе бактериальных геномов разного размера и разных таксономических групп.

Результаты. Созданные инструменты для классификации ТФ, кодируемых в геноме, и аннотации сайтов связывания реализованы в виде конвейера, принимающего на вход геномный файл в формате GenBank и выполняющего ряд последовательных этапов его обработки.

Инструмент для классификации ТФ выполняет средствами Biopython 4 отбор всех последовательностей, размеченных в исходном геноме как кодирующие, генерируя файл с открытыми рамками считывания, транслированными в аминокислотную последовательность в формате FASTA. Далее этот файл обрабатывается инструментом hmmscan 5 с использованием библиотеки, описывающей типовые семейства ДНК-связывающих доменов. Результаты hmmscan фильтруются, для каждого ТФ определяется последовательность CR-тега на основании координат модели ДНК-связывающего домена, и выгружаются в формате JSON для последующей отправки и отображения на клиентской части веб-приложения (рисунок 1).

а)

Annotation | TF classifier

TFs encoded in the Pve32_2022.07.gbk

Protein ID	Gene	Locus Tag	Critical Residue Tag	Family	Accession	E-value	Score	Description
AVT56699.1	asnC	OA04_00020	YSPGTH	HTH_AsnC-type	PF13404.7	1.7e-15	49.3	DNA-binding transcriptional dual regulator
AVT56705.1	qseB	OA04_00080	VSNAEVHHH	Trans_reg_C	PF00486.26	2e-23	75.0	two-component system response regulator
AVT56712.1	rbsR	OA04_00150	VATVSINSDTLNQ	HTH_LacI	SM00354	7.1e-29	92.3	transcriptional repressor RbsR
AVT56723.1	ntrC	OA04_00330	GRNTTR	bEBP_DBD	YN006	1.6e-18	58.9	nitrogen regulation protein NR(I)
AVT56741.1	-	OA04_00530	SPTHQPSQSR	HTH_18	PF12833.8	1.7e-22	72.2	AraC family transcriptional regulator
AVT56752.1	-	OA04_00640	-	Sigma70_r2	PF04542.16	3.2e-13	42.0	ECF family RNA polymerase sigma factor
AVT56763.1	relB1	OA04_00750	VSKKLR	PhdYeFM_antitox	PF02604.18	5.2e-08	25.3	antitoxin
AVT56769.1	-	OA04_00810	LASLSTKRPA	XRE_superfamily	SM00530	5.1e-08	25.5	XRE family transcriptional regulator, putative antitoxin component of HipA like toxin-antitoxin module
AVT56770.1	-	OA04_00820	YTSQSSSYR	HTH_1	PF00126.25	2.8e-17	55.1	LysR family transcriptional regulator

Total TF count: 339 — Download as TSV file

б)

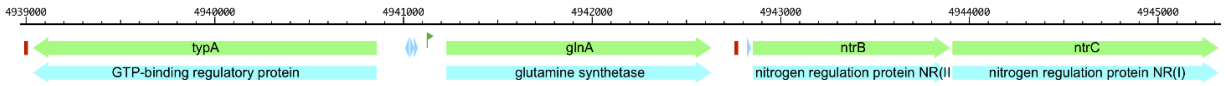

Рисунок 1 — Результаты классификации транскрипционных факторов (а) для генома Pectobacterium versatile 3-2 и аннотации (б) сайтов связывания NtrC. Сайты связывания добавлены с помощью инструмента аннотации и визуализированы (голубые треугольники) в программе SigmoID 6 (в приведенном фрагменте генома видно 4 из 22 аннотированных сайтов).

Инструмент для аннотации сайтов связывания помимо геномного файла принимает на вход информацию о профиле регуляторного мотива, который необходимо использовать. На базе входных данных выполняется сканирование с выбранной моделью посредством nhmmsearch 5 и последующая фильтрация результатов на основании пороговых значений модели. Идентифицированные сайты аннотируются в формате GenBank и предоставляются пользователю через интерфейс клиентской части веб-приложения.

Инструменты обработки бактериальных геномов реализованы в рамках веб-приложения, обеспечивающего удаленную работу пользователей с базой данных бактериальных регуляторных последовательностей BacRegDB [2, 3]. Разработанное веб-приложение включает серверную часть, представляющую собой RESTful Spring Boot приложение, написанное на Java и клиентскую часть, разработанную с использованием фреймворка ReactJS 7. Обмен данными между серверной и клиентской частями приложения выполняется в формате JSON с использованием http запросов, что обеспечивает возможность относительно независимой разработки отдельных частей приложения. Построение клиентской части в виде набо-

ра React компонент упрощает возможность расширения в перспективе перечня инструментов для анализа данных.

Пользовательский интерфейс вышеуказанного приложения включает страницу, позволяющую выбрать для обработки файл в формате GenBank, содержащий базовую разметку кодирующих последовательностей. При использовании инструмента для аннотации сайтов связывания необходимо выбрать ТФ из доступных в базе данных BacRegDB. Для упрощения поиска требуемого ТФ реализована фильтрация по семейству транскрипционных факторов и/или по их присутствию в загруженном пользователем файле. Непосредственная обработка предоставленного пользователем генома выполняется серверной частью веб-приложения. Результаты работы инструментов аннотации и классификации отображаются средствами разработанного веб-интерфейса в текстовом и табличном виде, соответственно. Дополнительно пользователю доступны для скачивания файл в формате GenBank с результатами аннотации, а также .tsv файл с результатами работы классификатора.

Разработанный набор средств анализа регуляторной информации был использован для классификации ТФ в геноме *Pectobacterium versatile* 3-2 и последующей аннотации сайтов связывания NtrC. Полученные результаты представлены на рисунке 1.

Заключение. Разработаны инструменты для классификации бактериальных транскрипционных факторов и аннотации сайтов связывания выбранного фактора в бактериальных геномах. Отличительной чертой предложенных инструментов является использование CR-тегов для ассоциации известных регуляторных мотивов с закодированными в геноме ТФ, что позволяет переносить регуляторную информацию в соответствии с более строгим критерием, чем уровень гомологии последовательностей ДНК-связывающих доменов. Программная реализация разработанных инструментов в рамках веб-приложения в сочетании с базой данных BacRegDB значительно повышает их доступность для практического использования. Созданный инструментарий применим для работы с геномами бактерий, имеющих базовую разметку кодирующих последовательностей и доступен по адресу http://bacregdb.bsu.by/tools.

Список литературы

1. Nikolaichik, Y. New approach to genome-wide automated inference of bacterial transcription factor binding sites / Y. Nikolaichik, P. Vychik // Abstracts of the XII Intern. Multiconf. "Bioinformatics of Genome Regulation and Structure/Systems Biology". – Novosibirsk, 2020. – P. 75–76.

2. Скакун, В. В. Разработка базы данных мотивов регуляции транскрипции у бактерий / В. В. Скакун, Е. А. Николайчик // Информатика 19 (1). 2022. – С. 59–71.

3. Дигрис, А. В. Веб-приложение для доступа к базе данных мотивов регуляции транскрипции у бактерий / А. В. Дигрис, Е. И. Дувалов, В. В. Скакун, Е. А. Николайчик. // Компьютерные технологии и анализ данных (CTDA’2020): материалы III Междунар. науч.-практ. конф., Минск, 21–22 апр. 2022 г. / Белорус. гос. ун-т ; редкол.: В. В. Скакун (отв. ред.) [и др.]. – Минск : РИВШ, 2022. С. 273-276.

4. Cock, P. J. et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics / P. J. Cock [et al.] // Bioinformatics. – 2009. – Vol. 25(11). – P.1422–1423.

5. Finn, R.D. HMMER web server: interactive sequence similarity searching / R. D. Finn, J. Clements, S. R. Eddy // Nucleic Acids Research. – 2011. – Vol. 39, № suppl. – P. W29-W37.

6. Nikolaichik, Y. SigmoID: a user-friendly tool for improving bacterial genome annotation through analysis of transcription control signals / Y. Nikolaichik, A.U. Damienikan // PeerJ. – 2016. – Vol. 4: e2056.

7. Порселло, Е. React: современные шаблоны для разработки приложений / Е. Порселло, А. Бэнкс. – 2^{e} изд. – Питер. – 2021. – 320 с.