ИДЕНТИФИКАЦИЯ ПРОМОТОРОВ НА ОСНОВЕ АНАЛИЗА СТРУКТУР АЛЬТЕРНАТИВНЫХ СИГМА-ФАКТОРОВ БАКТЕРИЙ

Д. И. Громыко, П. В. Вычик, Е. А. Николайчик

Белорусский Государственный Университет, Минск, Беларусь
E-mail: grom.dima.grom@gmail.com

Идентификация бактериальных промоторов (в отличие от эукариотических) до сих пор не имеет надежного алгоритмического решения, что во многом определяется уникальными особенностями распознавания промоторов сигма-факторами РНК-полимераз. В настоящей работе мы приводим анализ проблематики и предлагаем варианты решения этой задачи для промоторов, распознаваемых альтернативных сигма-факторами, на основе анализа доступных 3D-структур транскрипционных инициаторных комплексов.

Ключевые слова: *сигма-фактор, промотор, ДНК-мотив, ДНК связывающий домен.*

Развитие геномных технологий и растущие объемы генерируемой ими геномной информации требуют использования автоматических систем аннотации геномов. Такие системы (в основном программные конвейеры) разработаны и являются стандартом большинства геномных проектов. Однако имеющиеся конвейеры аннотации бактериальных геномов свою задачу выполняют лишь частично, идентифицируя преимущественно открытые рамки считывания, гены рРНК, тРНК и некоторых регуляторных РНК. Ни один из современных конвейеров не аннотирует регуляторные элементы (промоторы, операторы, терминаторы и др.), без чего использование автоматически аннотированных геномных последовательностей для многих целей ограничено.

Разработанная нами ранее программа SigmoID призвана закрыть этот пробел, однако ее первая версия могла аннотировать только терминаторы и известные операторы [1]. Во второй версии мы добавили конвейер для идентификации неизвестных операторов [2]. Конвейер эксплуатирует идею CR-тегов – последовательностей критичных, т. е. непосредственно контактирующих с азотистыми основаниями ДНК, аминокислотных остатков транскрипционных факторов. CR-теги можно вычислить путем анализа атомных координат комплексов транскрипционных факторов с ДНК, а идентификация соответствующего операторного мотива возможна за счет обычно имеющих место авторегуляции и/или сцепления гена транскрипционного фактора с хотя бы одним геном его регулона [2].

До настоящей работы мы не применяли такой подход к идентификации бактериальных промоторов из-за существенных отличий механизма

их распознавания. Промоторы распознаются сигма-факторами РНК-полимеразы, имеющими в своем составе два ДНК-связывающих домена, SR2 (взаимодействует с областью -10 промотора) и SR4 (взаимодействует с областью -35 промотора). SR4 контактирует с двухцепочечной ДНК 4–6 остатками одной α-спирали, как и большинство других ДНК-связывающих доменов. Однако SR2 отвечает за разделение цепей ДНК при инициации транскрипции и специфические контакты формирует в основном с одиночными (обеими) цепями ДНК, из-за чего общее число специфически взаимодействующих с ДНК аминокислотных остатков этого домена превышает три десятка. Суммарная длина полного CR-тега в результате может достигать 40 остатков, что делает наш конвейер неэффективным из-за слишком высокой специфичности отбора гомологичных сигма-факторов и недостаточного разнообразия соответствующих регуляторных областей.

Детальный анализ наиболее важных контактов в имеющихся структурах трех семейств транскрипционных факторов (таблица 1) с помощью алгоритма 3D-Footprint [3], а также выявление взаимокоррелирующих аминокислотно-нуклеотидных пар с помощью алгоритма Prot-DNA-Korr [4] позволило сократить CR-теги сигма-факторов до 13-19 остатков и успешно применить наш алгоритм для анализа промоторов нескольких семейств альтернативных сигма-факторов.

Таблица 1

КО-теги моделей альтернативных σ-факторов доступных в конвейере *de novo* поиска SigmoID

Сигма-фактор	**Семейство**	**Модель[а]**	**Проанализированные структуры[б]**	**Мотив найден[в]**
RpoN	RpoN	PF04552	5ui5,2o8k,2o9l,5nsr,5nss,6gfw,6gh5,6gh6	-
FliA	FliA_WhiG	TIGR02479	6pmi, 6pmj	+
RpoE	ECF	ECF02	2h27,4lup,2map,5or5,6jbq	+
HrpL	ECF	ECF32		+
FecI	ECF	ECF243		-

Примечания

[а] Использованы скрытые марковские модели из баз данных PFAM, TIGRFAMs и ECFhub

[б] Приведены коды доступа Protein Data Bank

[в] Идентификация мотива с помощью de novo конвейера SigmoID версии 2.0

Для проверки нашего подхода были выбраны пять сигма-факторов *Escherichia coli*. Критериями отбора служили наличие 3D-структур инициаторных комплексов, охарактеризованных промоторов и принадлежность к разным семействам. Пять выбранных сигма-факторов относились к семействам с существенными отличиями механизмов распознавания

ДНК: для ECF-факторов характерны типичные описанные выше контакты между промотором и доменами SR2 и SR4, FliA имеет вставку домена SR3, также способного контактировать с промотором, а в случае RpoN оба ДНК-связывающих домена взаимодействуют α-спиральными участками с двухцепочечной ДНК.

Применение модифицированного конвейера для идентификации промоторных мотивов для указанных сигма-факторов оказалось успешным для RpoN, FliA, RpoE и HrpL: найденные мотивы (рис. 1) обладали высоким сходством с описанными в литературе [5, 6]. Неудача с FecI оказалась вызвана очень малым разнообразием промоторных фрагментов (фактически с их идентичностью) для сигма-факторов с CR-тегом FecI.

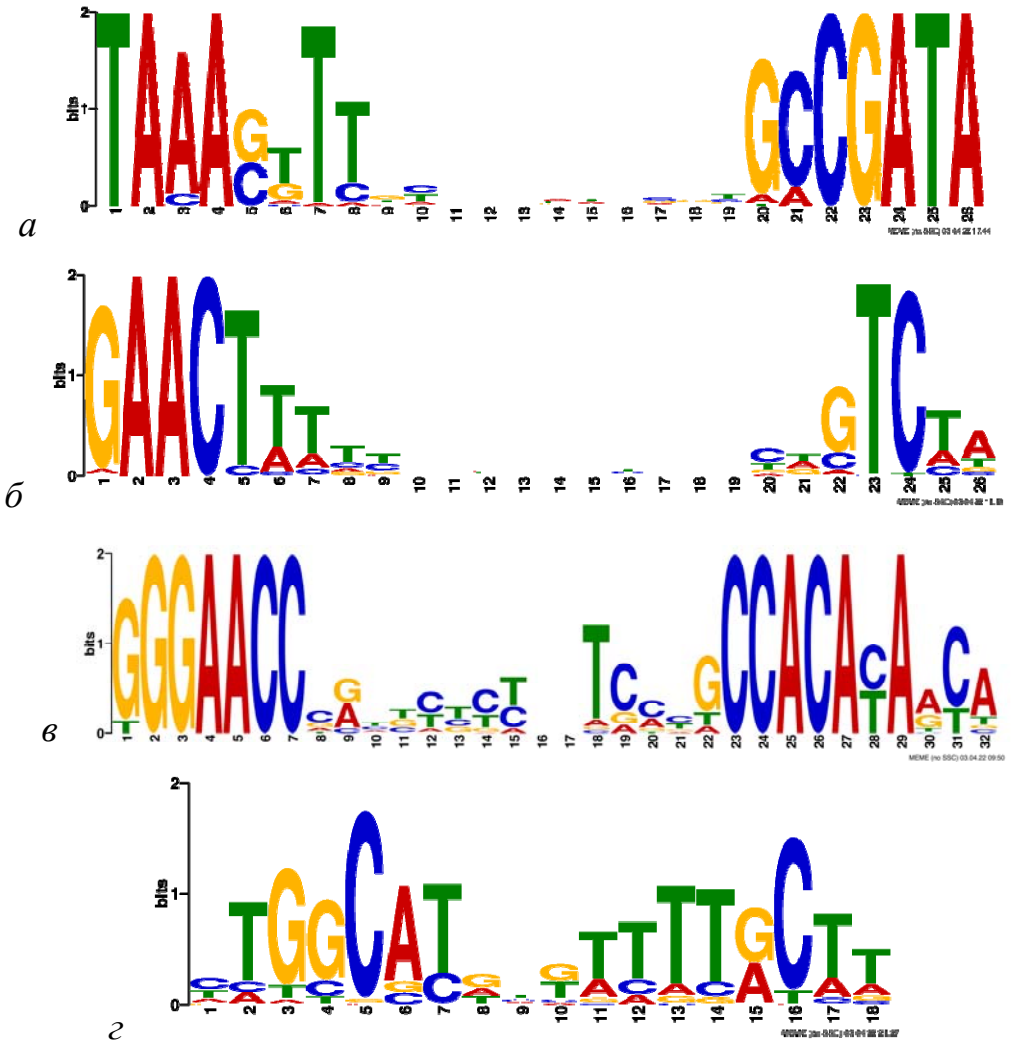

Рис. 1. Лого промоторных мотивов для сигма-факторов FliA (*а*), RpoE (*б*), HrpL (*в*) и RpoT (*г*)

Заключение. Выполненные в настоящей работе анализ 3D-структур инициаторных комплексов РНК-полимераз с их промоторами, идентификация критичных для распознавания промоторов аминокислотных остатков и модификация программного конвейера идентификации регуляторных последовательностей программы SigmoID показывают принципиальную возможность идентификации промоторных мотивов с помощью основанного на CR-тегах подхода, аналогично идентификации операторных последовательностей. Модифицированный код программы SigmoID и калиброванные профили для идентификации промоторов четырех альтернативных сигма-факторов доступны в репозитории github.com/nikolaichik/sigmoid. Калиброванные профили пригодны для автоматической аннотации промоторов в последовательностях бактериальных геномов.

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. 2016. Vol. 4, P. E2056.
2. Nikolaichik Y., Vychik P. New approach to genome-wide automated inference of bacterial transcription factor binding sites // Bioinformatics of Genome Regulation and Structure/ Systems Biology. V.21 N. 567. 2020. P. 75–76.
3. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein–DNA complexes // Nucleic Acids Research. 2010. Vol. 38, № Suppl_1. P. D91–D97.
4. Korostelev,Y.D. et al. Identification of Position-Specific Correlations between DNA-Binding Domains and Their Binding Sites. Application to the MerR Family of Transcription Factors // PLOS ONE. 2016. Vol. 11, № 9. P. E0162681.
5. Lam H.N., Chakravarthy S., Wei H.-L. et al. Global Analysis of the HrpL Regulon in the Plant Pathogen Pseudomonas syringae pv. tomato DC3000 Reveals New Regulon Members with Diverse Functions // PLOS ONE. 2014. V. 9. N. 8. P. e1061115.
6. Fitzgerald D. M., Smith C., Lapierre P., Wade J. T. The evolutionary impact of intragenic FliA promoters in proteobacteria // Mol. Microbiol. 2018. V. 108. P. 361–378.