ИДЕНТИФИКАЦИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ КАК ИНСТРУМЕНТ ИССЛЕДОВАНИЯ РЕГУЛЯТОРНЫХ СЕТЕЙ И РАЦИОНАЛЬНОГО ДИЗАЙНА БАКТЕРИАЛЬНЫХ ШТАММОВ

Николайчик Е.А., Вычик П.В.
Белорусский государственный университет, Минск, Беларусь
nikolaichik@bio.bsu.by

Конструирование бактериальных штаммов с заданными свойствами существенно облегчается пониманием устройства и механизмов функционирования регуляторных сетей, в основе которых лежит специфическое взаимодействие транскрипционных факторов с сайтами связывания в ДНК (операторами). Идентификация транскрипционного фактора, контролирующего определенный метаболический или биосинтетический путь, позволяет легко добиться повышенной экспрессии искомой группы генов путем инактивации или сверхэкспрессии гена транскрипционного фактора. С другой стороны, четкое определение операторного мотива позволяет выявить регулон, контролируемый конкретным транскрипционным фактором.

Экспериментальные исследования транскрипционной регуляции очень трудоемки и потому немногочисленны для немодельных организмов. В настоящей работе описаны основы применения универсального метода идентификации бактериальных операторов *in silico*, базирующегося на использовании информации об определяющих специфичность распознавания контактах аминокислотных остатков ДНК-связывающих доменов с азотистыми основаниями ДНК, и приведен пример анализа операторов в геноме *Pectobacterium atrosepticum*.

Анализ 3D-структур комплексов транскрипционных факторов с операторами позволяет определить критичные аминокислотные остатки, непосредственно контактирующие с ДНК и обеспечивающие ее специфическое распознавание. Позиции критичных остатков в ДНК-связывающих доменах конкретного белкового семейства строго консервативны. Скрытые марковские модели ДНК-связывающих доменов, доступные в базе данных PFAM и аналогичных ресурсах, позволяют идентифицировать гены почти всех транскрипционных факторов в конкретном бактериальном геноме. Поиск в белковых базах данных обычно дает обширный список гомологов конкретного транскрипционного фактора, из которого после отбора по критичным аминокислотным остаткам чаще всего остается от десятка до сотни гомологов с идентичными критичными остатками. Поскольку для большинства бактериальных транскрипционных факторов характерна авторегуляция, а также регуляция транскрипционных единиц, расположенных в геноме рядом с геном самого транскрипционного фактора, операторный мотив, распознаваемый

транскрипционным фактором в ДНК, можно идентифицировать за счет анализа регуляторных областей гена транскрипционного фактора и расположенных по обе стороны от него транскрипционных единиц, извлеченных из геномов, кодирующих гомологичные транскрипционные факторы того же семейства с идентичными критичными остатками.

Идея такого подхода впервые предложена в работе [4], а наша реализация улучшенного варианта алгоритма доступна в версии 2 программного пакета с открытым кодом SigmoID [5]. Программная реализация метода выполнена в среде Xojo с добавлениями на языке python. Открытый код и дистрибутивы программы доступны через репозиторий GitHub (github.com/nikolaichik/SigmoID). Тестирование метода выполнялось с использованием расшифрованных ранее [2] геномных последовательностей штаммов *Pectobacterium atrosepticum* 21A и *P. carotovorum* 3-2 (коды доступа GenBank CP009125 и CP024842).

Сканирование белковых последовательностей *P. atrosepticum* с помощью скрытых марковских моделей из базы данных PFAM выявило 262 транскрипционных фактора, принадлежащих к 28 семействам. Калиброванные профили операторных последовательностей созданы для каждого транскрипционного фактора из 10 семейств, представленных в исследуемом протеоме более чем пятью белками. В результате анализа 10 семейств транскрипционных факторов найдены наиболее вероятные операторы для 125 транскрипционных факторов (таблица). По результатам сравнения найденных мотивов с известными (опубликованными в литературе или присутствующими в базах данных прокариотических мотивов) метод оказался достаточно надежен, когда для исследуемого семейства транскрипционных факторов известны структуры хотя бы 4-5 не слишком близких ДНК-белковых комплексов. Однако даже в непростых случаях описываемый здесь подход может быть полезен. Например, транскрипционные факторы из LysR-семейства являются самыми распространенными у прокариот, но имеют довольно сложную структуру и плохо кристаллизуются в комплексе с ДНК, а их сайты связывания очень слабо охарактеризованы. Тем не менее, применение описанного здесь подхода на основе единственной присутствующей в Protein Data Bank структуры позволяет предсказать вероятные сайты связывания примерно для 2/3 транскрипционных факторов семейства LysR.

В настоящее время идет экспериментальная верификация наиболее интересных регуляторных взаимодействий, предсказанных для пектобактерий с помощью описанного здесь метода для трех основных семейств транскрипционных факторов: LysR, GntR и LuxR. Получены первые результаты, подтверждающие возможность конструирования штаммов-продуцентов на основе анализа транскрипционной регуляции *in silico* с использованием описанного здесь метода.

Таким образом, для типичной энтеробактерии удалось идентифицировать наиболее вероятные операторные мотивы (и регулоны) для примерно половины транскрипционных факторов. Эти мотивы требуют экспериментальной верификации, однако значительная их часть уже основана на экспериментальных данных, собранных в базах данных типа RegulonDB [6], CollecTF [1] и Prodoric [3], а транскрипционные факторы с идентичными контактами с ДНК встречаются и у эволюционно отдаленных организмов, что позволяет надежно "ретранслировать" подтвержденную информацию об операторах на организмы, для которых такой информации не имеется.

Таблица – Сайты связывания транскрипционных факторов в геноме *P. atrosepticum* 21A.

Семейство ТФ	модель PFAM/ SMART	Число ТФ с известными структурами	число ТФ в геноме		
			всего (известных)	доступны для анализа	идентифици-ровано корректных мотивов
LacI	PF00356	7	21 (14)	18	17 (94%)
GntR	PF00392	3	22 (9)	18	14 (78%)
LuxR	PF00196	4	12 (4)	12	8 (75%)
TetR	PF00440	7	17 (3)	16	13 (81%)
LysR	PF00126	1	59 (2)	50	32 (64%)
bEBP	PF02954	4	8 (5)	8	8 (100%)
OmpR	PF00486	5	13 (4)	13	10 (77%)
XRE	PF01381	11	19 (1)	15	12(80%)
HxlR	PF01638	1	8 (1)	8	6(75%)
MarR	SM00347	5	6 (0)	6	5(83%)

Работа выполнена при поддержке гранта БРФФИ Б18Р-117.

Литература

1. From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF / S. Kılıç [et al.] // Database J. Biol. Databases Curation. – 2016. – Vol. 2016 – P. baw055.

2. Genome Sequence of Pectobacterium atrosepticum Strain 21A / Y. Nikolaichik [et al.] // Genome Announc. – 2014. – Vol. 2, № 5. – P. e00935-14.

3. PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes / A. Grote [et al.] // Nucleic Acids Res. – 2009. – Vol. 37, № Database issue. – P. D61-D65.

4. Sahota, G. Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes / G. Sahota, G.D. Stormo // Bioinformatics. – 2010. – Vol. 26, № 21. – P. 2672-2677.

5. 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 – P. e2056.

6. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond / S. Gama-Castro [et al.] // Nucleic Acids Res. – 2016. – Vol. 44, № D1. – P. D133-D143.