СРАВНИТЕЛЬНЫЙ АНАЛИЗ САЙТОВ СВЯЗЫВАНИЯ БЕЛКОВ СЕМЕЙСТВ LuxR И LacI С ПОМОЩЬЮ НОВОГО АЛГОРИТМА

Вычик П.В., Николайчик Е.А.
Белорусский государственный университет, Минск,
p.vychik@gmail.com

Растущее число секвенированных бактериальных геномов открывает новые перспективы для прикладного использования данных организмов в биотехнологии и генной инженерии. Однако основным препятствием для практического использования этих объектов остается низкое качество аннотации геномных последовательностей и особенно отсутствие аннотации регуляторных элементов генома, без которых ничего нельзя сказать об экспрессии конкретного гена в определенных условиях.

Анализ сайтов связывания транскрипционных факторов (ССТФ) в новых геномах может осуществляться с использованием трудоемких и дорогостоящих экспериментальных методик (EMSA, SELEX, ДНК-футпринтинг), однако использование накопленного массива экспериментальных данных и методов биоинформатики позволяет существенно упростить процесс, уменьшая число потенциальных мишеней для экспериментальной проверки. Во многих случаях такой анализ *in silico* позволяет предложить простые способы повышения (или понижения) уровня экспрессии целевого гена.

Алгоритм поиска ССТФ *de novo*, добавленный в версии 2 разработанной ранее программы SigmoID [1], использует информацию о структурах кристаллизованных комплексов транскрипционных факторов (ТФ) со своими операторными последовательностями. Для представителей одного семейства ТФ критичные аминокислотные остатки ДНК-связывающего домена, специфически контактирующие с азотистыми основаниями ДНК, имеют фиксированные позиции и представляют собой своеобразный фингерпринт (КО-тэг), строго определяющий распознаваемую операторную последовательность [2].

Конвейер поиска операторных мотивов программы SigmoID v2 (рис. 1) использует библиотеку идентификаторов ТФ, полученную путем кластеризации ТФ на основании совпадения КО-тэгов, для экстракции геномных последовательностей из базы данных GenBank. Поскольку для большинства бактериальных ТФ характерна авторегуляция и, соответственно, расположение их операторов рядом с кодирующим ТФ геном, экстрагированные геномные последовательности содержат регуляторные области гомологичных ТФ с идентичными КО-тэгами из разных геномов. Идентификация операторных мотивов в полученном массиве регуляторных последовательностей осуществляется с помощью программы MEME.

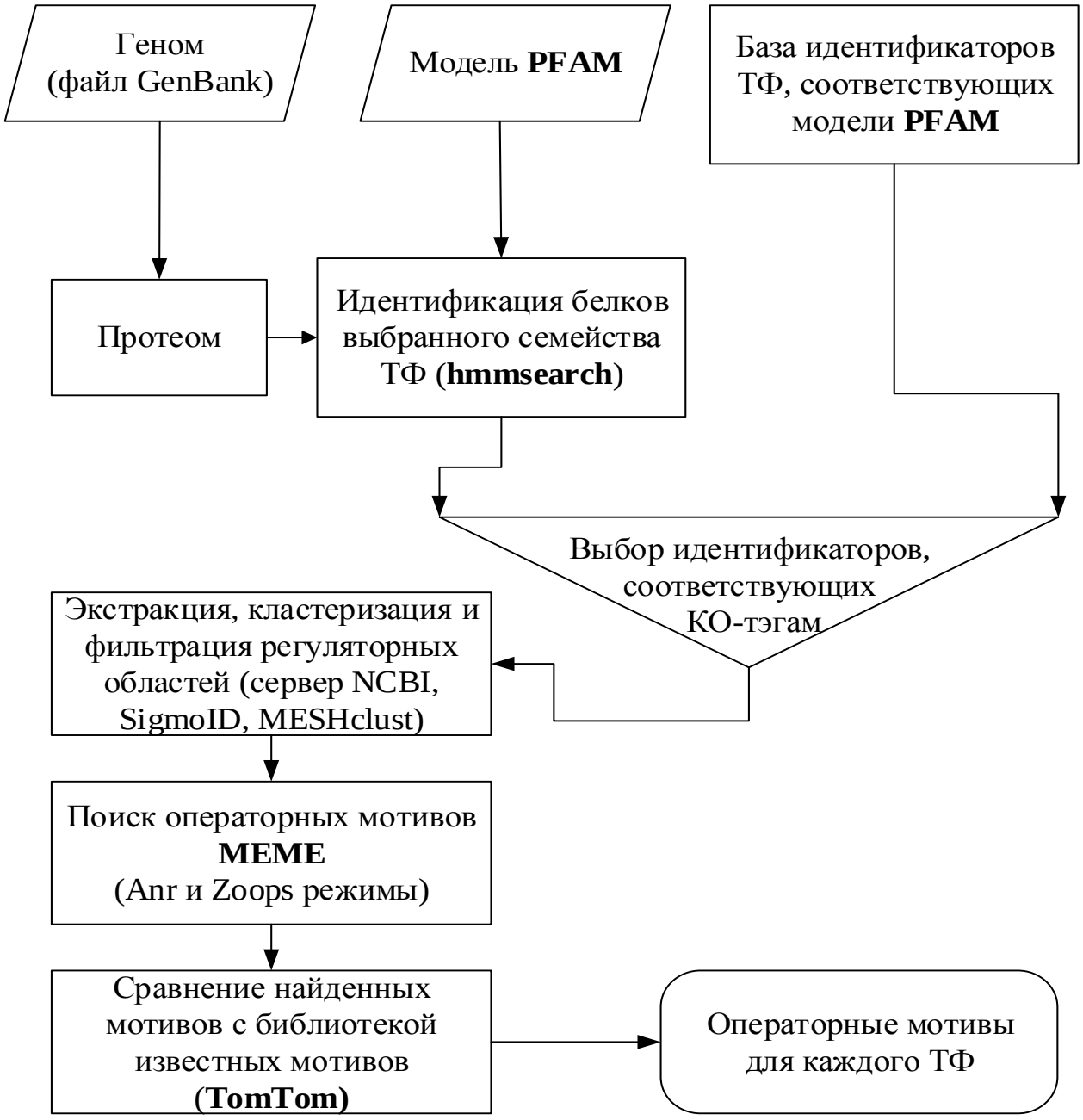

Рисунок 1 — Схема работы конвейера поиска ССТФ *de novo* программы SigmoID.

В настоящей работе эффективность нового алгоритма оценена путем анализа двух наиболее показательных семейств ТФ *Escherichia coli* K12 MG1655: LacI, для представителей которого полностью соблюдаются допущения метода (авторегуляция и гомодимерная структура ТФ) и LuxR, представители которого часто работают как гетеродимеры чьи гены часто удалены от генов-мишеней.

Для 11 из 12 ТФ семейства LacI по крайней мере один корректный мотив был найден (табл. 1). Проблемы с идентификацией мотива для CytR можно

объяснить необходимостью образования комплекса с CRP для связывания с ДНК.

Таблица 1 — Результаты поиска мотивов для ТФ семейства LacI.

Локус	ТФ	КО-тэг	Корректный мотив найден	Авто-регуляция	Регуляторных областей до/после кластеризации и фильтрации
NP_414879	LacI	LYSYQSRSHV	+	+	282/19
NP_415836	YcjW	IYSKSSRTNI	+	+	297/21
NP_417194	AscG	MLSKASRGYV	+	+	299/26
NP_416656	GalS	IRSVASRTLV	+	+	296/4
NP_417314	GalR	IKSVASRPKA	+	+	298/23
NP_418209	RbsR	MKSTSSHRFV	+	+	295/28
NP_418662	TreR	IKGKSSRSGV	+	+	297/21
NP_418685	IdnR	LQTKMSRKKV	+	+	300/21
NP_418369	CytR	MKSTASRDKV	-	+	297/20
NP_416175	PurR	IKSTTSHRFV	+	+	299/37
NP_416137	MalI	IHSVSSLGRI	+	+	295/16
YP_026222	GntR	LQTKMSREQV	+	+	299/27

Примечание: информация о наличии авторегуляции по данным RegulonDB [3].

Только для 5 из 19 ТФ семейства LuxR были найдены операторные мотивы (табл. 2). Этот результат может быть объяснен тем, что многие ТФ этого семейства образуют гетеродимеры с другими белками (RcsB, BglJ, GadE) [3], а также отсутствием авторегуляции, связанной с быстрой миграцией генов ТФ этого семейства путем горизонтального переноса, что часто разрывает связь между геномными позициями гена ТФ и его мишеней.

Представленный подход продемонстрировал высокую эффективность идентификации мотивов для ТФ семейства LacI — точность около 92%. Поиск мотивов ТФ семейства LuxR показал меньшую эффективность, только для 37% имеющихся в геноме ТФ были найдены корректные мотивы. Следует отметить, что даже в этом случае идентификация ТФ с учетом КО-тэгов может быть эффективной при использовании информации об операторах, экспериментально охарактеризованных у других видов бактерий, поскольку КО-тэги являются универсальной характеристикой ТФ, не имеющей видовых границ. Версия 2 программы SigmoID имеет большую библиотеку калиброванных скрытых марковских моделей экспериментально установленных операторов, "привязанных" к КО-тэгу (64 модели только для

семейства LuxR), которые можно использовать с любыми бактериальными геномами.

Таблица 2 — Результаты поиска мотивов для ТФ семейства LuxR.

Локус	ТФ	КО-тэг	Корректный мотив найден	Авто-регуляция	Регуляторных областей до/после кластеризации и фильтрации
NP_414828	MatA	KTYTHR	+	+	100/17
NP_417323	YqeH	RSYAYQ	-	нет данных	29/18
NP_417969	GadE	QTKIQF	-	+	35/26
NP_417964	DctR	KTYCHH	-	нет данных	112/20
NP_418785	YjjQ	KTSAQN	-	-	25/25
NP_414849	PdeL	KTSHQK	-	+	51/39
NP_418786	BglJ	KTRAHF	-	-	100/17
NP_416721	RcsB	KTSSQK	-	-	252/47
NP_416461	RcsA	KTSSHG	+	+	193/34
NP_417877	MalT	TTKTHR	+	+	258/72
NP_415558	CsgD	NTKTHY	-	+	124/38
NP_417977	YhjB	GTKAHE	-	нет данных	131/17
NP_416424	UvrY	KTNSYY	-	-	223/23
NP_416870	EvgA	KTSTYS	+	+	246/66
NP_415739	NarL	STKVHK	+	+	235/47
NP_416426	SdiA	NTNFHK	-	-	200/39
NP_418125	UhpA	KTHVHA	+	+	238/45
NP_415068	FimZ	KTSAHS	-	нет данных	55/18
NP_416697	NarP	QTKVHR	+	+	79/14

Примечание: информация о наличии авторегуляции по данным RegulonDB [3].

Литература

1. 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.

2. 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.

3. 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.