ИДЕНТИФИКАЦИЯ И КЛАССИФИКАЦИЯ БАКТЕРИАЛЬНЫХ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

Е. А. Николайчик, П. В. Вычик

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

Разработан метод автоматической идентификации и классификации бактериальных ДНК-связывающих транскрипционных факторов. Метод корректно определяет принадлежность транскрипционных факторов к 74 семействам и трем суперсемействам и превосходит по своей чувствительности и селективности имеющиеся решения. Метод реализован в рамках программы анализа транскрипционной регуляции бактерий SigmoID и является частью автоматизированного конвейера аннотации операторных элементов в бактериальных геномных последовательностях. Открытый код разработки и исполняемые файлы доступны в репозитории github.com/nikolaichik/sigmoid.

Ключевые слова: *транскрипционный фактор, оператор, аннотация бактериальных геномов.*

Существенным недостатком депонированных в базах данных последовательностей бактериальных геномов является неполная аннотация регуляторного компонента, включая гены транскрипционных факторов и сайты их связывания с ДНК (операторы и промоторы). В большинстве случаев имеется только общая аннотация части транскрипционных факторов без детализации их функций или хотя бы принадлежности к определенному семейству, многие транскрипционные факторы в принципе не аннотированы как таковые, а аннотация распознаваемых ими операторов и промоторов почти всегда отсутствует.

Отсутствие аннотации регуляторных элементов в геномных последовательностях затрудняет работу с соответствующими штаммами, в особенности все, что связано с адаптацией бактерии к меняющимся условиям (в том числе в ходе колонизации эукариотических организмов-хозяев). Актуальность регуляторной информации подчеркивают наблюдения о том, что в транскриптомных экспериментах с достаточной глубиной покрытия не удается детектировать экспрессию до половины бактериальных генов. Полная аннотация регуляторного компонента генома позволит понять, в каких условиях может экспрессироваться большинство генов. Такое понимание актуально не только для фундаментальных исследований, но и в практических целях (например, при конструировании штаммов-продуцентов для биотехнологических производств).

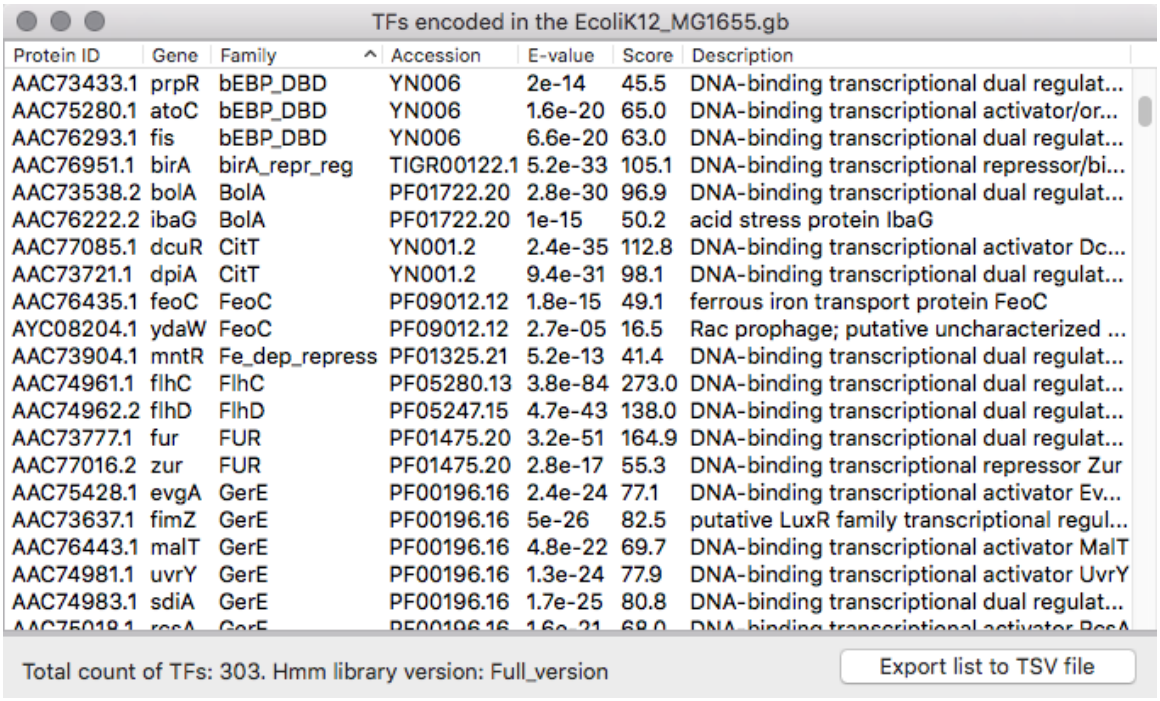
Разработанный нами ранее программный пакет SigmoID предназначен для идентификации и аннотации регуляторных последовательностей в бактериальных геномах [1]. В версии 2 SigmoID добавлены возможности

анализа и аннотации регуляторных последовательностей экспериментально неохарактеризованных транскрипционных факторов [2], однако такой анализ выполнялся только для избранных семейств транскрипционных факторов, представленных в библиотеке программы. В настоящей работе эти возможности дополнены полноценным модулем для идентификации, классификации и аннотации транскрипционных факторов, что минимизирует вероятность пропуска важных транскрипционных регуляторов и максимально приближает анализ регуляторной информации к полногеномному.

Основу классификатора составляет библиотека скрытых марковских моделей ДНК-связывающих доменов бактериальных транскрипционных факторов. По возможности использованы калиброванные модели из баз данных PFAM, SMART и TIGRFAMs [3–5]. Отбор моделей производился путем сканирования коллекции изученных транскрипционных факторов из баз данных RegulonDB, CollecTF, Prodoric2, CoryneRegNet, DBTBS [6–10]. Дополнительно транскрипционные факторы были идентифицированы в протеомах модельных штаммов бактерий с помощью алгоритма глубокого обучения DeepTFactor [11] с последующей детальной верификацией каждого нового транскрипционного фактора. Из коллекции транскрипционных факторов удалялись белки без ДНК-связывающего домена, нуклеоид-связывающие белки, рекомбиназы и метилазы. При отсутствии подходящей модели выполнялись поиск гомологов исследуемых транскрипционных факторов, множественное выравнивание с опорой на имеющиеся 3D-структуры с помощью алгоритма T-Coffee Expresso [12] и последующим редактированием и построением соответствующей скрытой марковской модели. Для трех семейств (MarR, XRE и MerR) на основании тщательного анализа удалось снизить пороговые битовые значения для идентификации с помощью этих моделей транскрипционных факторов соответствующих суперсемейств и сокращения числа необходимых моделей. Полученная таким образом HMM-библиотека имеет в своем составе 77 моделей (64 PFAM, 4 SMART, 2 TIGR, 7 сконструированных в этой работе) и способна идентифицировать большинство известных бактериальных транскрипционных факторов.

Для идентификации транскрипционных факторов с помощью полученных скрытых марковских моделей применяется программа hmmscan из пакета HMMER [13], а принадлежность к определенному семейству определяется по критерию e-value результатов поиска hmmsearch с моделью соответствующего ДНК-связывающего домена. Пользовательский интерфейс и вывод результатов классификации в табличном формате (рис. 1) реализованы в среде разработки Xojo. Соответствующие функ-

ции включены в открытый код программы SigmoID и доступны из репозитория github.com/nikolaichik/sigmoid.


TFs encoded in the EcoliK12_MG1655.gb

Protein ID	Gene	Family	Accession	E-value	Score	Description
AAC73433.1	prpR	bEBP_DBD	YN006	2e-14	45.5	DNA-binding transcriptional dual regulat...
AAC75280.1	atoC	bEBP_DBD	YN006	1.6e-20	65.0	DNA-binding transcriptional activator/or...
AAC76293.1	fis	bEBP_DBD	YN006	6.6e-20	63.0	DNA-binding transcriptional dual regulat...
AAC76951.1	birA	birA_repr_reg	TIGR00122.1	5.2e-33	105.1	DNA-binding transcriptional repressor/bi...
AAC73538.2	bolA	BolA	PF01722.20	2.8e-30	96.9	DNA-binding transcriptional dual regulat...
AAC76222.2	ibaG	BolA	PF01722.20	1e-15	50.2	acid stress protein IbaG
AAC77085.1	dcuR	CitT	YN001.2	2.4e-35	112.8	DNA-binding transcriptional activator Dc...
AAC73721.1	dpiA	CitT	YN001.2	9.4e-31	98.1	DNA-binding transcriptional dual regulat...
AAC76435.1	feoC	FeoC	PF09012.12	1.8e-15	49.1	ferrous iron transport protein FeoC
AYC08204.1	ydaW	FeoC	PF09012.12	2.7e-05	16.5	Rac prophage; putative uncharacterized ...
AAC73904.1	mntR	Fe_dep_repress	PF01325.21	5.2e-13	41.4	DNA-binding transcriptional dual regulat...
AAC74961.1	flhC	FlhC	PF05280.13	3.8e-84	273.0	DNA-binding transcriptional dual regulat...
AAC74962.2	flhD	FlhD	PF05247.15	4.7e-43	138.0	DNA-binding transcriptional dual regulat...
AAC73777.1	fur	FUR	PF01475.20	3.2e-51	164.9	DNA-binding transcriptional dual regulat...
AAC77016.2	zur	FUR	PF01475.20	2.8e-17	55.3	DNA-binding transcriptional repressor Zur
AAC75428.1	evgA	GerE	PF00196.16	2.4e-24	77.1	DNA-binding transcriptional activator Ev...
AAC73637.1	fimZ	GerE	PF00196.16	5e-26	82.5	putative LuxR family transcriptional regul...
AAC76443.1	malT	GerE	PF00196.16	4.8e-22	69.7	DNA-binding transcriptional activator MalT
AAC74981.1	uvrY	GerE	PF00196.16	1.3e-24	77.9	DNA-binding transcriptional activator UvrY
AAC74983.1	sdiA	GerE	PF00196.16	1.7e-25	80.8	DNA-binding transcriptional dual regulat...

Total count of TFs: 303. Hmm library version: Full_version

Export list to TSV file

Рис. 1. Табличное представление результата работы классификатора транскрипционных факторов в программе SigmoID

Сравнение с имеющими аналогичный функционал ресурсами показывает большую чувствительность разработанного классификатора. Так, для наиболее изученного модельного организма *Escherichia coli* известный ресурс P2TF [14] определяет 273 транскрипционных фактора, тогда как наш классификатор – 303. (валидация результатов проводилась на основе доступных в литературе экспериментальных данных).

Для 28 семейств (включая три суперсемейства) транскрипционных факторов, имеющих известные структуры комплексов фактора с оператором с помощью ресурса Interaction service портала NPIDB [15] идентифицированы аминокислотные остатки, специфически распознающие азотистые основания нуклеотидов в составе операторов, что делает их пригодными для непосредственного использования в конвейере идентификации операторных последовательностей de novo версии 2 программы SigmoID.

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. DOI: 10.7717/peerj.2056
2. Nikolaichik Y., Vychik P. Genome-wide inference of bacterial transcription factor binding sites: new method and its applications // BMC Bioinformatics. 2020. Vol. 21, № S20. P. O2. DOI: 10.1186/s12859-020-03838-2
3. Finn R.D. et al. The Pfam protein families database: towards a more sustainable future // Nucleic Acids Research. 2016. Vol. 44, № D1. P. D279–D285. DOI: 10.1093/nar/gkv1344
4. Letunic I., Bork P. 20 years of the SMART protein domain annotation resource // Nucleic Acids Research. 2018. Vol. 46, № D1. P. D493–D496. DOI: 10.1093/nar/gkx922
5. Haft D.H. et al. TIGRFAMs and Genome Properties in 2013 // Nucleic Acids Research. 2012. Vol. 41, № D1. P. D387–D395. DOI: 10.1093/nar/gks1234
6. Santos-Zavaleta A. et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12 // Nucleic Acids Research. 2019. Vol. 47, № D1. P. D212–D220. DOI: 10.1093/nar/gky1077
7. Kılıç S. et al. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria // Nucleic Acids Research. 2013. Vol. 42, № D1. P. D156–D160. DOI: 10.1093/nar/gkt1123
8. Dudek C.-A., Jahn, D. PRODORIC: state-of-the-art database of prokaryotic gene regulation // Nucleic Acids Research. 2022. Vol. 50, № D1. P. D295–D302. DOI: 10.1093/nar/gkab1110
9. Parise M.T.D. et al. CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks // Scientific Data. 2020. Vol. 7, № 1. P. 142. DOI: 10.1038/s41597-020-0484-9
10. Sierro N. et al. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information // Nucleic Acids Research. 2008. Vol. 36, № Suppl_1. P. D93–D96. DOI: 10.1093/nar/gkm910
11. Kim G.B. et al. DeepTFactor: A deep learning-based tool for the prediction of transcription factors // Proceedings of the National Academy of Sciences. 2021. Vol. 118, № 2. P. E2021171118. DOI: 10.1073/pnas.2021171118
12. Armougom F. et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee // Nucleic Acids Research. 2006. Vol. 34, № Web server. P. W604–W608. DOI: 10.1093/nar/gkl092
13. Finn R.D., Clements J., Eddy S.R. HMMER web server: interactive sequence similarity searching // Nucleic Acids Research. – 2011. Vol. 39, № Suppl 2. P. W29–W37. DOI: 10.1093/nar/gkr367
14. Ortet P. et al. P2TF: a comprehensive resource for analysis of prokaryotic transcription factors // BMC Genomics. 2012. Vol. 13, № 1. P. 628. DOI: 10.1186/1471-2164-13-628