doi 10.18699/bgrs2024-1.2-04

Вычислительный конвейер по распознаванию *de novo* сайтов связывания транскрипционных факторов в бактериальных геномах

Мухин А.М.[1, 2, 3]*, Ощепков Д.Ю.[1, 2]

[1] *Институт цитологии и генетики СО РАН, Новосибирск, Россия*

[2] *Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия*

[3] *Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия*

* *mukhin@bionet.nsc.ru*

Ключевые слова: Поиск *de novo* мотивов, филогенетический футпринтинг, вычислительный конвейер, Python, Nextflow

Мотивация и цель: Аннотация бактериальных геномов и их конкретных регуляторных геномных последовательностей сайтами связывания транскрипционных факторов (ССТФ) является актуальной биологической задачей, поскольку работа клетки и производство определенных ферментов критическим образом зависит от существующих регуляторных геномных связей. Существующие программные решения содержат необходимые программные компоненты лишь частично, не позволяя комплексно и с минимальными трудозатратами выполнять массовый поиск ССТФ как во вновь секвенированных, так и недостаточно изученных бактериальных геномах. В данной работе реализован вычислительный конвейер, состоящий из следующих необходимых этапов для полноценной аннотации бактериальных геномов с помощью известных подходов *de novo* поиска ССТФ: (1) аннотирование генома оперонной структурой, необходимое для дальнейшего точного определения регуляторных/промоторных областей, (2) поиск *de novo* мотивов в промоторах целевого генома, (3) функциональной аннотации вновь выявленных ССТФ. Поиск *de novo* мотивов в промоторах целевого генома может осуществляться альтернативно с помощью двух подходов: либо в полной выборке промоторов целевого организма, либо на основании подхода филогенетического футпринтинга, осуществляя поиск ССТФ в наборе промоторов ортологичных генов из одной таксономической группы с целевым организмом. В последнем случае необходимый список этапов аннотации должен включать также (4) инструмент для поиска ортологичных генов заданного таксономического уровня и (5) базы данных с последовательностями и аннотациями известных бактериальных геномов, что позволит автоматически осуществлять все необходимые операции по формированию требуемых выборок промоторов. Такое комплексное решение подразумевает также наличие всех необходимых программных модулей, осуществляющих формирование необходимых выборок, операции по конвертации форматов и перенаправлению данных и сохранения необходимых данных в служебных БД. Такой комплексный подход позволит сократить до минимума затраты ресурсов на промежуточные, но требующие квалификации в программировании для персонала с другой (микробиологической) квалификацией.

Методы и алгоритмы: Оба альтернативных конвейера поиска *de novo* мотивов - по полной выборке промоторов целевого организма и с помощью подхода филогенетического футпринтинга были реализованы с помощью следующих инструментов: язык программирования Python, база данных SQLite и PostgreSQL, системы конвейеризации NextFlow.
Результаты: Реализован конвейер частично с использованием платформы nextflow и набора скриптов на языках программирования Python и bash. На этапе по определению оперонной структуры используется альтернативно либо БД DOOR 2.0 [1] для известных организмов, либо веб-сервис Operon Mapper [2] для вновь секвенированных геномов. Для поиска *de novo* мотивов по полной выборке промоторов целевого организма используется инструмент BoBro2 [3], подход филогенетического футпринтинга реализован внутри программного модуля MP3, объединяющего значительное количество известных подходов для выявления ССТФ de novo в бактериальных геномах [4]. Также для построения выборок промоторов ортологичных генов используется инструмент для поиска ортологов внутри заданной таксономической группы на основе их белковых последовательностей GOST [5]. Необходимый этап функциональной аннотации вновь выявленных ССТФ осуществляется с помощью приложения Tomtom программного пакета MEME suite [6] путем сравнения выявленных мотивов с известными мотивами из соответствующих БД. Для повышения скорости расчетов была развернута база данных SQLite использующаяся для индексации координат генов, оперонов, данных по таксономической принадлежности геномов и пр.
Выводы: Создан вычислительный конвейер по распознаванию *de novo* сайтов связывания транскрипционных факторов в бактериальных геномах на основе двух подходов: по полной выборке промоторов целевого организма, либо на основании подхода филогенетического футпринтинга. Конвейер включает необходимый набор инструментов для осуществления всех необходимых промежуточных вычислений, который можно использовать как в локальной среде, так на кластере. В дальнейшем планируется интеграция дополнительного модуля по определению оперонной структуры с использованием алгоритма машинного или глубокого обучения; включение в программный комплекс других известных подходов для поиска ССТФ de novo, интеграцию дополнительных средств отображения, обработки и ранжирования получаемых данных; расширение интегрированной базы данных для сохранения результатов расчетов, требующих значительных вычислительных мощностей; замена СУБД на PostgreSQL.
Финансирование: Исследование поддержано в рамках Программы Курчатовского геномного центра ИЦиГ СОРАН (№ 075-15-2019-1662).

A computational pipeline for *de novo* recognition of transcription factor binding sites in bacterial genomes

Mukhin A.M. [1, 2, 3]*, Oshchepkov D.Y. [1, 2]

[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*

[2] *Kurchatov Genomic Center ICiG SB RAS, Novosibirsk, Russia*

[3] *Novosibirsk National Research State University, Novosibirsk, Russia*

* *mukhin@bionet.nsc.ru*

Ключевые слова: de novo motifs, phylogenetic footprinting, computational pipeline, Python, Nextflow

Motivation and aim: Annotation of bacterial genomes and their specific regulatory genomic sequences by transcription factor binding sites (TFBS) is an urgent biological task, since cell function and production of certain enzymes critically depend on existing regulatory genomic connections. Existing software solutions contain the necessary software components only partially, not allowing a comprehensive and labor-intensive mass search for TFBS in both newly sequenced and insufficiently studied bacterial genomes. In this work, we have implemented a computational pipeline consisting of the following necessary steps for full annotation of bacterial genomes using known *de novo* search approaches for TFBS: (1) annotation of the genome with operon structure necessary to further pinpoint regulatory/promoter regions, (2) *de novo* search for motifs in the promoters of the target genome, and (3) functional annotation of newly identified TFBS. The *de novo* motifs search in the promoters of the target genome can be performed alternatively using two approaches: either in the full set of promoters of the target organism, or based on a phylogenetic footprinting approach, searching for TFBS in a set of promoters of orthologous genes from the same taxonomic group as the target organism. In the latter case, the required list of annotation steps should also include (4) a tool to search for orthologous genes of a given taxonomic level and (5) databases with sequences and annotations of known bacterial genomes, allowing to automatically perform all necessary operations to generate the required promoter samples. Such a complex solution also implies the presence of all necessary program modules, which perform the formation of the required samples, operations on format conversion and data redirection and saving the required data in the service database. Such an integrated approach will minimize the resources spent on intermediate, but requiring programming skills for personnel with other (microbiological) qualifications.
Methods and algorithms: Both alternative pipelines for *de novo* motif search - by full sampling of target organism promoters and by phylogenetic footprinting approach - were implemented using the following tools: Python programming language, SQLite and PostgreSQL database, NextFlow pipelining system.
Results: The pipeline is implemented partially using the nextflow platform and a set of scripts in Python and bash programming languages. The operon structure determination step alternatively uses either the DOOR 2.0 database [1] for known organisms or the web service Operon Mapper [2] for newly sequenced genomes. The BoBro2 tool [3] is used to search for de novo motifs from a full sample of promoters of the target organism; the phylogenetic footprinting approach is implemented within the MP3 software module, which integrates a significant number of known approaches for detecting *de novo* TFBS in bacterial genomes [4]. Also, a tool GOST for searching orthologous genes within a given taxonomic group based on their protein sequences is used to construct orthologous gene promoter samples [5]. The necessary step of functional annotation of newly identified TFBS is performed using the Tomtom application of the MEME suite software [6] by comparing the identified motifs with known motifs from the corresponding databases. To increase the speed of calculations, we deployed a SQLite database used for indexing the coordinates of genes, operons, data on the taxonomic affiliation of genomes, etc.
Conclusions: A computational pipeline for de novo recognition of transcription factor binding sites in bacterial genomes has been developed based on two approaches: full sampling of target organism promoters or phylogenetic footprinting. The pipeline includes the necessary set of tools to perform all necessary intermediate calculations, which can be used both in a local environment and on a cluster. The future plans include

the following steps: (a) to integrate an additional module for operon structure determination using a machine learning or deep learning algorithm; (b) to include other known approaches for *de novo* search of TFBS into the program complex; (c) to integrate additional tools for displaying, processing, and ranking the obtained data; (d) to expand the integrated database for saving the results of calculations that require significant computing power; (e) to replace the DBMS with PostgreSQL.

Funding: The research was supported under the Kurchatov Genomic Center Program of ICG SB RAN (No. 075-15-2019-1662).

Список литературы/References

1. Mao X. et al. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res*. 2014;42(D1):D654-D659
2. Taboada B. et al. Operon-mapper: A web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*. 2018;34(23):4118-4120
3. Ma Q. et al. An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*. 2013;29(18):2261-2268
4. Liu B. et al. An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes. *BMC Genomics.* 2016;17:578
5. Li G. et al. Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. *Nucleic Acids Res*. 2011;39(22):e150
6. Bailey T.L. et al. The MEME Suite. *Nucleic Acids Res*. 2015;43(W1):W39-W49