

Идентификация потенциальных промоторных последовательностей, предсказанных методом MADHS, путем анализа доступности хроматина и поиска неаннотированных сайтов старта транскрипции

Бубнова А.*, Гайдукова С., Яковлева И., Камионская А.

ФГУ «Федеральный исследовательский центр «Фундаментальные основы биотехнологии»
Российской академии наук», Москва, Россия

* an_bubnova@mail.ru

Ключевые слова: предсказанный потенциальный промотор; сайт начала транскрипции; доступная область хроматина

Мотивация и цель: Поиск, предсказание и идентификация неаннотированных промоторных последовательностей может иметь фундаментальное значение для обнаружения в геномах функциональных областей, не аннотированных ранее. Но на данный момент существующие алгоритмы прогнозирования промоторов не исключают появления ложноположительных результатов, что является серьезной проблемой вычислительной биологии. В связи с этим при наличии предсказанных промоторных последовательностей целесообразным является проведение дополнительной проверки их свойств для подтверждения возможности их функционирования в геноме в роли регуляторных элементов. Целью данной работы была проверка предсказанных методом MADHS [1] промоторных последовательностей 1-й хромосомы генома риса *O. sativa* на наличие потенциальных неаннотированных сайтов старта транскрипции в районе 3'-конца предсказанных промоторов, а также изучение доступности хроматина в районе предсказанных последовательностей.

Методы и алгоритмы: Предсказанные методом MADHS промоторные последовательности были выгружены из Базы данных потенциальных промоторных последовательностей (доступно по адресу: <http://victoria.biengi.ac.ru/cgi-bin/dbPPS/index.cgi>), всего было отобрано 126 последовательностей из хромосомы 1 генома риса с наибольшим аргументом нормального распределения, который показывает степень неслучайности созданного выравнивания при предсказании. В геноме эти последовательности расположены на расстоянии не ближе 1000 п. н. от известных генов и не ближе, чем на 200 п. н. от транспозонов и SINE повторов. Это гарантирует, что предсказанные потенциальные промоторы не связаны с транспозонами и промоторами известных генов. Длина выбранных предсказанных потенциальных промоторов варьирует от 428 до 599 нуклеотидов, а расстояние от них до ближайших аннотированных транскриптов – от 1044 до 92 434 п. н.

Поиск транскрибируемых участков ДНК (потенциальных неаннотированных сайтов старта транскрипции) за потенциальными промоторными последовательностями проводился путем анализа результатов секвенирования кэп-анализов экспрессии генов (CAGE-seq) риса *O. sativa*, доступных в SRA NCBI. Данный анализ позволяет картировать большинство сайтов начала транскрипции

и, как следствие, их промоторов. Адаптеры и чтения низкого качества в необработанных данных секвенирования удалялись с помощью Trim-Galore. Выравнивание отфильтрованных данных на геном первой хромосомы риса проводилось с помощью *bowtie2* с опцией *very-sensitive*. Полученный в результате bam-файл был конвертирован в sam-файл, проиндексирован и отсортирован с помощью инструментов Samtools, затем из этого файла были получены пики, характеризующие участки потенциальных сайтов старта транскрипции. Расстояние от них до предсказанных промоторов оценивали с использованием Bedtools closest. В результате отбирались предсказанные промоторы, за которыми в направлении 3' были обнаружены потенциальные сайты начала транскрипции в пределах 1000 п. н. Визуализация выбранных данных осуществлялась с помощью PyGenomeTracks.

Структура хроматина анализировалась путем изучения данных секвенирования ATAC-seq, также взятых из SRA NCBI, позволяющих распознать открытые участки хроматина, доступные для посадки факторов транскрипции. Такие участки могут помогать в идентификации промоторных последовательностей. Алгоритм действий с данными ATAC-seq был идентичен вышеописанной работе с данными CAGE-seq, за исключением последнего этапа: после выявления пиков, характеризующих открытую область хроматина, отбирались промоторы, имеющие пересечение с пиками, тогда как расстояние до ближайших пиков не изучалось.

Результаты: При анализе данных CAGE-seq было обнаружено наличие четких пиков, характерных для областей сайтов старта транскрипции, расположенных в свободных от аннотированных транскриптов участках генома. Это может свидетельствовать о возможном существовании ранее неаннотированных транскриптов в геноме риса. Достоверность полученных результатов была подтверждена наличием пиков CAGE-seq в районах 5'-концов аннотированных транскриптов. Из 126 отобранных ранее предсказанных промоторных последовательностей было выявлено 14 промоторов, имеющих за собой потенциальные сайты старта транскрипции:

- 3 потенциальных TSS расположены не далее 100 п.н. от 3'-концов предсказанных промоторов;
- 9 потенциальных TSS, пересекающихся с 3'-концами предсказанных промоторов;
- 2 потенциальных TSS, расположенных на расстоянии от 100 до 1000 п. н. от 3'-концов предсказанных промоторов.

В результате анализа ATAC-seq из 126 предсказанных промоторов было отобрано 16 последовательностей, пересекающихся с пиками ATAC-seq. Это указывает на то, что предсказанные промоторы расположены в открытых участках хроматина, доступных для посадки транскрипционных факторов.

Выводы: Из 126 первоначально отобранных предсказанных промоторных последовательностей: 14 предсказанных промоторов имеют за собой потенциальную транскрипцию и 16 находятся в открытой области хроматина. Только 4 предсказанных промотора удовлетворяют обоим условиям проверки. Данные последовательности наиболее вероятно являются промоторными областями и рекомендованы к дальнейшей экспериментальной проверке *in vivo*.

Финансирование: Исследование выполнено при поддержке Минобрнауки России в рамках соглашения № 075-15-2022-318 от 20.04.2022 о предоставлении гранта в

форме субсидий из федерального бюджета на осуществление государственной поддержки создания и развития научного центра мирового уровня «Агротехнологии будущего».

Identification of potential promoter sequences predicted by the MADHS method by analyzing chromatin accessibility and searching for unannotated transcription start sites

Bubnova A.*, Gaidukova S., Yakovleva I., Kamionskaya A.

Federal State Institution "Federal Research Centre "Fundamentals of Biotechnology" of the Russian Academy of Sciences", Moscow, Russia

* an_bubnova@mail.ru

Key words: potential predicted promotor; transcription start site; chromatin accessibility

Motivation and Aim: Search, prediction, and identification of unannotated promoter sequences can play a fundamental role in the discovery of functional regions in genomes that have not been previously annotated. However, currently existing promoter prediction algorithms do not exclude the occurrence of false positives, which is a serious problem in computational biology. Therefore, in the presence of predicted promoter sequences, it is reasonable to carry out additional testing of their properties to confirm the possibility of their functioning in the genome as regulatory elements. The aim of this work was to check the predicted by the MADHS method [1] promoter sequences of the 1st chromosome of the rice *O. sativa* genome for the presence of potential unannotated transcription start sites near their 3'-end, as well as to study chromatin accessibility in the region of the predicted sequences.

Methods and Algorithms: Promoter sequences predicted by the MADHS method were downloaded from the Potential Promoter Sequences Database (available at <http://victoria.biengi.ac.ru/cgi-bin/dbPPS/index.cgi>), and 126 sequences with the largest normal distribution argument were selected from chromosome 1 of the rice genome, indicating the degree of nonrandom alignment created during the prediction process. In the genome, these sequences are located no closer than 1,000 bp from known genes and no closer than 200 bp from transposons and SINE repeats. This ensures that the predicted potential promoters are not linked to transposons and promoters of known genes. The length of the selected predicted potential promoters ranges from 428 to 599 nucleotides, and the distance from them to the nearest annotated transcripts ranges from 1,044 to 92,434 bp.

A search for transcribed DNA regions (potential unannotated transcription start sites) behind potential promoter sequences was performed by analyzing the results of gene expression sequencing analysis (CAGE-seq) of rice *O. sativa* available at NCBI SRA. This analysis allows mapping of most transcription start sites and, as a consequence, their promoters. Adapters and low-quality reads in the raw sequencing data were removed using the Trim-Galore program. Alignment of the filtered data to the genome of the first rice chromosome was performed using the bowtie2 program with the -very-sensitive option. The resulting bam file was converted to a sam file, indexed and sorted using Samtools tools, then peaks characterizing regions of potential transcription start sites were extracted. The distance from these to predicted promoters was estimated using

Bedtools closest. As a result, predicted promoters were selected, behind which potential transcription start sites within 1000 bp were detected in the 3' direction. The selected data were visualized using PyGenomeTracks.

Chromatin structure was analyzed using ATAC-seq sequencing data, also taken from the NCBI SRA, which allows us to recognize open regions of chromatin available for landing transcription factors. Such regions can help in the identification of promoter sequences. The algorithm of actions with ATAC-seq data was identical to the work with CAGE-seq data described above, except for the last step: after identifying the peaks characterizing the open region of chromatin, promoters overlapping with the peaks were selected, and the distance to the nearest peaks was not studied.

Results: Analysis of CAGE-seq data revealed the presence of distinct peaks characteristic of transcription start site regions located in regions of the genome that do not contain annotated transcripts. This may indicate the possible existence of previously unannotated transcripts in the rice genome. The validity of the results was confirmed by the presence of CAGE-seq peaks in the 5'-terminal regions of annotated transcripts. Out of 126 selected previously predicted promoter sequences, 14 promoters with potential transcription start sites behind them were identified:

- 3 potential TSS located no further than 100 bp from the 3'-end of the predicted promoters;
- 9 potential TSSs overlapping with the 3'-ends of predicted pro-motors;
- 2 potential TSS located between 100 and 1000 bp from the 3'-end of predicted promoters.

ATAC-seq analysis selected 16 sequences that overlapped with ATAC-seq peaks from 126 predicted promoters. This indicates that the predicted promoters are located in open regions of chromatin accessible for seeding of transcription factors.

Conclusion: Of the 126 previously predicted promoter sequences selected: 14 predicted promoters have potential transcription behind them and 16 are located in the open region of chromatin. Only 4 predicted promoters satisfy both validation conditions. These sequences are the most likely promoter regions and are recommended for further experimental verification *in vivo*.

Funding: The study is supported by the Ministry of Science and Higher Education of the Russian Federation in accordance with agreement No. 075-15-2022-318 date April 20, 2022 on providing a grant in the form of subsidies from the Federal budget of Russian Federation. The grant was provided for state support for the creation and development of a World-class Scientific Center “Agrotechnologies for the Future”.

Список литературы/References

1. Korotkov E.V. et al. Mathematical Algorithm for Identification of Eukaryotic Promoter Sequences. *Symmetry*. 2021;13(6):917