DOI: 10.25205/978-5-4437-1843-9-34

ПРЕДСКАЗАНИЕ САЙТОВ ПОСАДКИ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ ПО АМИНОКИСЛОТНЫМ ПОСЛЕДОВАТЕЛЬНОСТЯМ БЕЛКОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES BY AMINO ACID SEQUENCES OF PROTEINS USING ARTIFICIAL INTELLIGENCE METHODS *

Н. А. Шелудяков[1], П. С. Деменков[2]

[1]*Новосибирский государственный университет*
[2]*Институт цитологии и генетики СО РАН, Новосибирск*

N. A. Sheludyakov[1], P. S. Demenkov[2]

[1]*Novosibirsk State University*
[2]*Institute of Cytology and Genetics SB RAS, Novosibirsk*

✉ n.sheludyakov@g.nsu.ru

Аннотация

В работе представлен прототип инструмента Prot2Motif, позволяющий по аминокислотной последовательности транскрипционного фактора предсказывать мотивы сайтов посадки на ДНК. Основа модели — рекуррентная нейронная сеть, позволяющая работать как с биологическими последовательностями, так и с их матричными представлениями.

Abstract

A prototype of the Prot2Motif tool is presented in this work. Prot2Motif enables the prediction of DNA binding site motifs based on the amino acid sequence of a transcription factor. The core of the model is a recurrent neural network, which can process both biological sequences and their matrix representations.

В последние десятилетия биоинформатика стала передовой областью научных исследований, которые предоставляют новые возможности для понимания генетической основы жизни. Одна из задач данной области — поиск сайтов посадки транскрипционных факторов (СПТФ). Транскрипционные факторы (ТФ) — это белки, контролирующие первый этап экспрессии генов. СПТФ — это короткие последовательности нуклеотидов длиной от 5 до 30 нуклеотидов, которые распознаются специальными ДНК-связывающими доменами ТФ. Поиск таких коротких последовательностей позволяет лучше понимать процесс регуляции экспрессии генов на уровне транскрипции.

Существует множество как экспериментальных методик, так и вычислительных инструментов, позволяющих решать данную задачу. Однако инструменты *in silico* могут предсказывать сайты только на основе априорной информации о нуклеотидной последовательности. Поэтому мы разработали прототип инструмента Prot2Motif, который может предсказать сайты посадки исходя из первичной структуры белка. Основа модели — рекуррентная нейронная сеть (RNN), позволяющая генерировать последовательности нуклеотидов.

Наш инструмент включает в себя несколько компонентов. Первым звеном является модель ESM-2 (Evolutionary Scale Modeling), которая переводит последовательность аминокислот в вектора (эмбеддинги). Второе звено — InterProScan, необходимый для поиска ДНК-связывающего домена в белке. Третий компонент — рекуррентная нейронная сеть, позволяющая генерировать частоты встречаемости нуклеотидов в позициях.

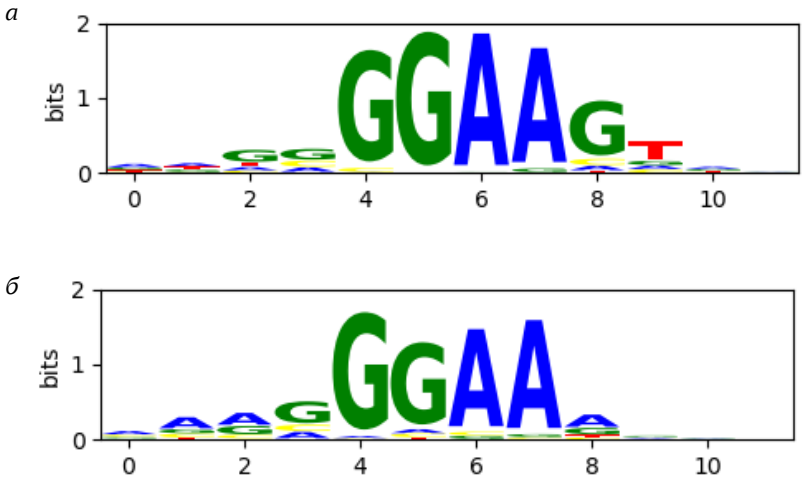
В качестве набора данных по мотивам сайтов связывания была взята база данных HOCOMOCO v13. Информация о первичной структуре белков была извлечена из баз данных UniProt и TFclass.

Для оценки качества нейронной сети была выбрана оценка SWscore [1], выражаемая формулой

$$\mathrm{SWscore} = \frac{\sum_{i=1}^{m}\left(2 - \sum_{j\in\{A,C,G,T\}}\left(A_{ij} - B_{ij}\right)^2\right)}{2m},$$

где m — длина мотива (m = 25); A_{ij} — вероятность нуклеотида j в позиции i исходной матрицы; B_{ij} — вероятность нуклеотида j в позиции i сгенерированной матрицы. Оценка SWscore может принимать значения от 0 до 1, где 1 — точное совпадение.

После обучения модели RNN качество предсказания на тестовой выборке составило 0,792, что свидетельствует о возможности модели находить взаимосвязи между структурой мотива сайта посадки и ДНК-связывающего домена белка. На рисунке представлен пример референсного и предсказанного мотива.



а — референсное LOGO-представление SPI1.H13CORE.1.S.B;
б — предсказанное LOGO-представление. Белок — Uniprot ID: P17947

В процессе обучения и оценки качества генерации замечено, что модель способна определять консервативность нуклеотидов в позициях — это выражается в корректном выборе наиболее часто встречаемого нуклеотида.

На сегодняшний день проводится оптимизация процесса обучения нейронной сети, что позволит улучшить качество генерации мотивов сайтов связывания.

Литература

1. Sandelin A., Wasserman W.W. Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics // Journal of Molecular Biology. 2004. Vol. 338, No. 2. P. 207–215.