Секция «Биоинформатика»

Объединение позиционно-весовых матриц в решающие деревья для распознавания сайтов связывания факторов транскрипции

Научный руководитель – Кулаковский Иван Владимирович

Кравченко Павел Андреевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: Pavel-Kravchenk0@yandex.ru

ДНК-паттерны, распознаваемые белками-регуляторами транскрипции (транскрипционными факторами, ТФ), традиционно представляются в виде позиционно-весовых матриц (ПВМ), которые предполагают независимость соседних нуклеотидов в сайтах связывания. В настоящее время предложено множество альтернативных моделей, учитывающих корреляции между соседними позициями или возможные подтипы паттернов, однако простые модели в виде ПВМ продолжают широко использоваться на практике.

Одним из способов построения уточненных моделей является объединение нескольких ПВМ в решающее дерево [1], при этом ранние подходы не показывали значительного прироста точности распознавания сайтов связывания ТФ, по сравнению с одной ПВМ. На сегодняшний день доступны результаты десятков независимых экспериментов для одного фактора транскрипции, и это позволяет построить множество ПВМ и применить современные методы машинного обучения, такие как градиентный бустинг, для построения объединенного классификатора [2].

В нашей работе мы использовали ПВМ, построенные по данным ChIP-Seq экспериментов, представленных в базе GTRD (Gene Transcription Regulation Database) [3] и ПВМ, полученные на их основе в ходе построения коллекции мотивов связывания ТФ мыши и человека HOCOMOCO [4]. Предсказания индивидуальных ПВМ использовались как признаки для обучения итоговой модели. В качестве позитивной выборке использовались участки связывания конкретного фактора транскрипции, определенные с помощью ChIP-Seq; в качестве негативной выборки использовались последовательности схожих длин, являющиеся сайтами связывания факторов транскрипции других структурных семейств.

Нам удалось продемонстрировать, что модель, построенная с использованием множества ПВМ, обладает улучшенной точностью предсказания сайтов для различных ТФ, причем эффект сохраняется как при предсказании сайтов связывания ТФ мыши с помощью моделей, обученных на данных ChIP-Seq человека, так и при обратной постановке. Было выяснено, что лучший результат достигается при объединении предсказаний мононуклеотидных и динуклеотидных ПВМ.

Реализация предложенного метода доступна в репозитории GitHub:
http://github.com/Pavel-Kravchenko/TF-ML

Источники и литература

1) Yingtao Bi, et al. Tree-Based Position Weight Matrix Approach to Model Transcription Factor Binding Site Profiles //PLoS One. 2011; 6(9)
2) Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System // arXiv. 2016

3) Yevshin I.S., et al. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments // Nucleic Acids Res. 2017 Jan 4; 45: D61–D67

4) Kulakovskiy I.V., et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // Nucleic Acids Res. 2018 Jan 4;46(D1):D252-D259