Секция «Биоинформатика»

Улучшение предсказания сайтов связывания транскрипционных факторов с помощью машинного обучения

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, однако ПВМ продолжают широко использоваться на практике.

В частности, одним из способов построения уточненных моделей является объединение нескольких ПВМ в решающее дерево [1], при этом ранее опубликованный подход не показал значительного прироста точности распознавания сайтов связывания ТФ, по сравнению с одной ПВМ. В то же время, на сегодняшний день наличие результатов десятков независимых экспериментов для одного фактора транскрипции позволяет построить множество ПВМ и затем 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 негативной выборки использовались последовательности схожих длин, являющиеся сайтами связывания факторов транскрипции других структурных семейств.

Нам удалось продемонстрировать, что модель, построенная с использованием множества ПВМ, позволяет значительно улучшить точность предсказания сайтов для различных ТФ, причем эффект сохраняется при предсказании сайтов связывания ТФ мыши при обучении на ChIP-Seq для ТФ человека и обратно.

Тестовая реализация доступна в репозитории GitHub:
http://github.com/Pavel-Kravchenko/TF-ML

Слова благодарности
Автор благодарит Пензара Д.Д., Воронцова И.Е. и Кулаковского И.В. за оказанную во время выполнения работы помощь.

Источники и литература

1 Yingtao Bi, et al. Tree-Based Position Weight Matrix Approach to Model Transcription Factor Binding Site Profiles. PLoS One. 2011.

2 Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. Carlos Guestrin University of Washington. 2016.

3 Yevshin I.S., et al. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic acids research. 2016.

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 research. 2018.