Секция «Молекулярная и клеточная биоинженерия и биоинформатика»

Использование методов машинного обучения для предсказания промоторных последовательностей *E. coli*

Научный руководитель – Сорокин Анатолий Александрович

Орлов М.А.[1], *Ермак Т.В.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Биологический факультет, Кафедра биофизики, Москва, Россия; 2 - Институт цитологии и генетики СО РАН, Новосибирск, Россия

Современные методы секвенирования предоставили возможность многократно ускорить и удешевить получение данных о первичной структуре ДНК, что привело к их огромному накоплению и необходимости использования методов автоматизированной аннотации. Решение данной задачи, особенно в случае новых геномов (тотальное и de novo секвенирование) может быть затруднено. Так, большинство используемых алгоритмов рассматривают нуклеотидную последовательность, что позволяет успешно предсказывать кодирующие участки генома, но не области с регуляторными функциями (в частности, промоторы). Для функционирования таких областей существенна не сама нуклеотидная последовательность, а кодируемые ей физические свойства, поскольку именно они определяют ДНК-белковые взаимодействия. Для аннотирования областей регуляции более перспективны алгоритмы, использующие профили физических свойств ДНК; кроме того, при этом целесообразно совместное использование различных некоррелирующих между собой физических, а также текстовых характеристик [1].

В данной работе для промоторов, "непромоторов" (отдалены на 300 п.о. и более от точки старта транскрипции), промоторных островков и генов E. coli (штамм K12) из базы данных RegulonDB версии 8.5 получены профили физических свойств ДНК: электростатического потенциала, энергии активации и размера открытых состояний ДНК (согласно модели [2]), а также рассчитанный скользящим окном GC-состав в качестве текстовой характеристики. Все профили рассчитаны для интервалов 200 нуклеотидов. С использованием этих данных, а также набора редуцированных профилей (полученных на основе всех 4 характеристик при помощи метода анализа главных компонент) была проведена кластеризация ("машинное обучение без учителя") и оценена точность бинарных классификаторов ("машинное обучения с учителем") для пары "промоторы — последовательности другого типа". Кластерный анализ выполнен по методу Уорда с предварительной оценкой оптимального числа кластеров (при помощи анализа их устойчивости и топологии дендрограмм). В результате для ряда кластеров описано наличие характерных элементов профилей и обогащение функциональными классами соответствующих им генов (согласно GeneOntology). Показано, что совместное использование физических свойств промоторных последовательностей при кластерном анализе позволяет более эффективно отличать их от последовательностей ДНК других типов. Бинарные классификаторы Naive Bayes и Random Forest продемонстрировали высокую точность предсказаний для всех пар "промоторы — последовательности другого типа", причем модели второй группы показали лучшие результаты. Работа поддержана грантом РФФИ №16-37-00303 мол_а.

Источники и литература

1) *A.A. Grinevich, A.A. Ryasik, L.V. Yakushevich* Trajectories of DNA bubbles // Chaos, Solitons and Fractals, 75, 62, 2015

2) *H.Q. Wang, C.J. Benham.* Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress // BMC Bioinformatics Vol. 7, 2006, pp. 248-262.