КЛАССИФИКАТОРЫ ГЕНОМА E.COLI НА ОСНОВЕ АНАЛИЗА ПРОФИЛЕЙ ФИЗИЧЕСКИХ ХАРАКТЕРИСТИК ДНК

Рясик А.А., Орлов М.А., Зыкова Е.А., Ермак Т.В.[1], Сорокин А.А.

Институт Биофизики Клетки РАН, Пущино, Россия
[1]Институт Цитологии и Генетики СО РАН, Новосибирск, Россия

В настоящее время скорость секвенирования геномов превышает скорость их аннотации. При этом ставший классическим подход в биоинформатике анализа только текстовых последовательностей не позволяет адекватным образом описывать ДНК-белковые взаимодействия, а, следовательно, определять положение регуляторных участков ДНК. Кроме того, явление молекулярной мимикрии показало, что белки не способны “читать” ДНК-последовательности. В данном случае роль в ДНК-белковом узнавании и взаимодействии играют физические характеристики молекулы ДНК (электростатический потенциал).

Поэтому естественным подходом для решения задачи о предсказании регуляторных областей является использование профилей распределения физических характеристик вдоль молекулы ДНК. Раннее было показано, что такой подход даёт лучшие результаты по сравнению с предсказаниями, сделанными только на основе текстового анализа [1]. Поэтому для увеличения точности предсказаний мы предложили использование сразу нескольких характеристик. Для этого мы выбрали характеристики различной природы, и обладающие различными математическими свойствами: динамические характеристики открытых состояний — их энергия активации и размер [2], распределение электростатического потенциала на поверхности ДНК и рассчитанный скользящим окном GC-состав.

На основе данных характеристик были построены парные классификаторы методами Naive Bayes и Random Forest. Для классификационного анализа были подготовлены пять наборов фрагментов ДНК: 699 экспериментально подтвержденных промоторов E. coli, гены E. coli, антипромоторы, промоторные островки и участки ДНК, находящиеся на расстоянии не менее 300 п.о. от известных промоторов (непромоторы). Анализ результатов обучения показал, что модели показывают одинаковые параметры точности, специфичности и чувствительности для всех групп — 100% кроме группы “промоторы против не промоторов”.

Работа поддержана грантом РФФИ №16-37-00303 мол_а.

1. Темлякова Е.А. “Роль электростатического потенциала ДНК в формировании промоторной функции в геноме E. coli” дисс. канд. физ.-мат. наук: 03.01.02, 12.05.2016 Пущино

2. A.A. Grinevich, A.A. Ryasik, L.V. Yakushevich, Trajectories of DNA bubbles Chaos, Solitons & Fractals, v. 75, p. 62, 2015;