РАЦИОНАЛЬНАЯ ОРГАНИЗАЦИЯ ФИЗИЧЕСКИХ ПАРАМЕТРОВ ДУПЛЕКСА ДНК С ЦЕЛЬЮ УЛУЧШЕНИЯ РАСПОЗНАВАНИЯ ПРОМОТОРНЫХ ОБЛАСТЕЙ

Орлов М.А., Рясик А.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

Увеличение числа *de novo* секвенированных геномов определяет необходимость развития методов их автоматизированной аннотации. Алгоритмы, анализирующие первичную структуру ДНК, позволяют успешно предсказывать положение кодирующих областей, однако для предсказания положения регуляторных областей генома (особенно промоторов) они малоэффективны. В настоящее время наиболее перспективными считаются алгоритмы, которые наряду с текстовыми используют физические характеристики ДНК, напрямую определяющие ход процесса ДНК-белкового взаимодействия. Причем наиболее эффективно использование нескольких таких свойств одновременно [1].

В данной работе для полного набора экспериментально подтвержденных промоторов *E. coli* (штамм K12) из базы данных RegulonDB версии 8.5 получены профили физических свойств, представляющих различные типы: электростатический потенциал, вызванная суперспирализацией дестабилизация дуплекса (SIDD, Stress-Induced Duplex Destabilization) и данные модели динамических свойств открытых состояний ДНК.

Для каждого из наборов профилей, а также набора, полученного с помощью PCA (88%), получены устойчивые кластеры, для которых установлено наличие характеристических элементов профилей и обогащение функциональными классами соответствующих генов (по GeneOnthology). Показано, что совместное использование редуцированных с помощью PCA наборов физических свойств промоторных последовательностей при кластерном анализе позволяет более эффективно отличать их от последовательностей ДНК других типов. Использованная организация данных может быть использована совместно с другими методами машинного обучения.Работа поддержана грантом РФФИ №16-37-00303 мол_а.

Литература.

1. Wang, H.Q. and Benham, C.J. 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.