ПРЕДСКАЗАНИЕ ПОЛОЖЕНИЯ САЙТОВ ПОСАДКИ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ E.COLI ПРИ ПОМОЩИ QSAM-ПРЕДСТАВЛЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Темлякова Е.А., Джелядин Т.Р., Сорокин А.А.

ИБК РАН, Россия, 142290, Пущино, ул.Институтская, 3, evgenia.teml@gmail.com

Регуляция экспрессии генов на уровне транскрипции является важной характеристикой метаболизма прокариотических организмов. Синтез РНК-транскрипта с молекулы ДНК начинается со специальных участков – промоторов, которые в свою очередь могут регулироваться посредством белков – транскрипционных факторов (ТФ). Поиск сайтов посадки транскрипционных факторов (СПТФ) – одна из важнейших задач при изучении регуляции транскрипции у прокариот. При этом возникает две принципиально разные подзадачи: (1) предсказание наличия регуляции того или иного промотора и (2) поиск точного положения сайта посадки в случае наличия регуляции. В большинстве алгоритмов, направленных на поиск СПТФ, используются позиционные весовые матрицы [1, 2, 3]. Однако такой подход приводит к возникновению большого числа ложно-положительных предсказаний при относительно низкой специфичности метода.

В нашей работе мы предлагаем анализировать не текстовую последовательность СПТФ, а распределение физико-химических свойств, задаваемых конкретной нуклеотидной последовательностью. При этом каждому нуклеотиду сопоставляется n значений в n-мерном пространстве его физико-химических характеристик. Такое представление нуклеотидных последовательностей получило название QSAM-представления (quantitative sequence-activity models, [4]) и является разновидностью QSAR-моделирования (quantitative structure–activity relationship), широко применяемого для предсказания биологической активности химических соединений. Используя QSAM-представления последовательностей СПТФ и методы машинного обучения с учителем, мы построили несколько типов классификаторов, обученных для определения потенциальных сайтов посадки. Показано, что при использовании QSAM-представления вместо исходной текстовой последовательности, несмотря на сходное количество ложно-положительных срабатываний, точность решения обеих задач, представленных выше, оказывается заметно выше.

Литература.

1. *Stormo G. D., et al* Characterization of translational initiation sites in *E. coli* // *NAR.* **10**, 9, 1982. Стр 2971-2996.
2. *Münch R. et al.* PRODORIC: prokaryotic database of gene regulation // *NAR.* **31**, 1, 2003. Стр 266-269.
3. *Salgado H. et al.* RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more // *NAR.* **41**, D1, 2013. Стр D203-D213.
4. *Sandberg M., et al* A multivariate characterization of tRNA nucleosides // *J. Chemometrics.* **10**, 5-6, 1996. Стр 493-508.