

**Универсальные нейросетевые архитектуры для предсказания сайтов
связывания факторов транскрипции**

Научный руководитель – Зинкевич Арсений Олегович

Бушуев Сергей Евгеньевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Москва, Россия

E-mail: bushsebz@gmail.com

Важным свойством функционирования клетки является регуляция транскрипции, основанная на специализированных белках — транскрипционных факторах, позволяющая клетке адаптироваться к различным условиям среды. Изучение и описание транскрипционных факторов, как и распознаваемых ими наборами нуклеотидных последовательностей — мотивов, представляет собой ценность для определения фенотипа на основании условий среды и генома.

В данной работе рассматривается создание нейронной сети, которая способна предсказывать связывание определенным транскрипционным фактором (ТФ) последовательностей, полученных при помощи нескольких различных методов: *in vitro* (HT-SELEX) и *in vivo* (ChIP-Seq и GHT-SELEX). Фундаментальной разницей этих групп методов является источник отбираемых последовательностей: первая группа методов основана на синтезированных последовательностях, вторая на геномной библиотеке.

В рамках такой задачи нашей целью было реализовать модель, способную к переносу выученных паттернов между этими методами, причем возможны обучение паттернам на результатах экспериментов *in vitro*, их применение на результатах экспериментов *in vivo* и обратное (A2G и G2A, ‘Artificial to Genomic’ и ‘Genomic to Artificial’, соответственно). Данные задачи были представлены основными номинациями прошедшего в этом году соревнования The IBIS Challenge, разработанного в нашей лаборатории, на основании которого была проведена наша работа.

Предсказание связывания нуклеотидной последовательности фактором транскрипции было формализовано как задача бинарной классификации. В качестве входных данных для обучения модели мы использовали последовательности, полученные методом HT-SELEX. Объектами положительного класса были отмечены последовательности, связываемые целевым белком, отрицательные контроли с подобным распределением GC-состава были случайно выбраны из последовательностей, связываемых другими белками. Модель была обучена предсказывать вероятность связывания белком определенной последовательности, полученной методом HT-SELEX, ChIP-Seq или GHT-SELEX. В качестве оптимизируемой функции была выбрана бинарная кросс-энтропия.

В качестве основы для архитектуры модели была использована нейронная сеть LegNet, имеющая EfficientNet-подобную архитектуру, занявшую 1-е место в предсказании регуляторной активности последовательностей в международном конкурсе DREAM-2022.

Мы модифицировали данную архитектуру, интегрировав технические и архитектурные особенности подходов других участников соревнования в сочетании с заменой задачи классификации на регрессию номера цикла HT-SELEX, формированием части негативной выборки путем перемешивания букв в последовательностях позитивной выборки и агрегацией предсказаний модели на центрированных регионах разных длин.

В результате предложенная нами модель занимает первое место в пост-конкурсном анализе. По оценкам текущих тестов, наибольший вклад в качество модели внесло одновременное использование сразу нескольких методов для формирования негативной выборки.