Секция «Биоинформатика»

Предсказание сигналов ядерной локализации при помощи машинного обучения

Научный руководитель – Залевский Артур Олегович

Романова Татьяна Андреевна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: tatiana.romanova@student.msu.ru

Правильная внутриклеточная локализация белка чрезвычайно важна для его функционирования. Для транспорта в ядро белки, большие 40 кДа, должны обладать сигналом ядерной локализации (NLS). В классическом случае NLS связывается с импортином-α, который взаимодействует с импортином-β, обеспечивающим прохождение через ядерные поры.

Большинство существующих систем предсказания NLS опираются преимущественно на поиск паттернов в последовательности, лишь некоторые также учитывают предсказаную неструктурированность региона или данные о белок-белковых взаимодействиях. Однако недавно AlphaFold[1] предоставил для множества белков предсказанные структуры достаточно высокого качества. Они позволяют точнее иных методов определять неструктурированность[2], а также учитывать другие характеристики, такие как общая экспонированность участка.

Из более 139 тыс. доступных AlphaFold структур белков 8 модельных эукариотических организмов (*Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* и др.) мы извлекли около 2000 аннотированных NLS. Для каждого мы получили набор функциональных и физико-химических декскрипторов: заряд, поверхность, доступная растворителю, разметка вторичной структуры и т.д. Поскольку имеющиеся в базах данные о взаимодействии белков с импортинами достаточно скудны, для оценки возможности взаимодействия использовались результаты молекулярного докинга фрагментов белка в репрезентативную выборку структур импортина-α.

При тестировании на экспериментально валидированных NLS модель на основе алгоритма случайного леса показала следующие результаты: precision = 0.68, recall = 0.61, F1-метрика = 0.64, поаминокислотная точность предсказания (aPC) = 0.49. По F1-метрике наша модель превосходит такой широко используемый алгоритм как NLStradamus на 51%, недавний (2020) метод INSP на 12%; aPC на тех же данных превышает таковую для NLStradamus и INSP более чем вдвое[3].

В текущий момент мы проводим поиск новых неаннотированных NLS в эукариотических белках, которые затем будут экспериментально проверены.

Благодарности:
Работа выполнена с использованием оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ имени М.В. Ломоносова.

Источники и литература

1) Jumper J. et al. Highly accurate protein structure prediction with AlphaFold: 7873 // Nature. Nature Publishing Group, 2021. Vol. 596, № 7873. P. 583–589.
2) Akdel M. et al. A structural biology community assessment of AlphaFold 2 applications. bioRxiv, 2021. P. 2021.09.26.461876.
3) Guo Y. et al. Discovering nuclear targeting signal sequence through protein language learning and multivariate analysis // Anal. Biochem. 2020. Vol. 591. P. 113565.