ПОСТРОЕНИЕ КЛАССИФИКАТОРА ДЛЯ ВЫЯВЛЕНИЯ БЕЛКОВ, РАЗДЕЛЯЮЩИХ ЖИДКИЕ ФАЗЫ, МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Грибкова А.К., Шайтан А.К.

Биологический факультет, Московский Государственный Университет им. М.В. Ломоносова, Россия, 119991, Москва, 1-12 Ленинские горы

В клетках живых организмов находятся спонтанно возникающие капли, не окруженные мембранами. Капли образуются в ходе процесса разделения жидких фаз (liquid-liquid phase separation, LLPS) и имеют значение для протекания ряда биохимических реакций. Несмотря на возросший интерес к этой теме, экспериментальные исследования остаются разрозненными, что является предпосылкой для комплексного биоинформатического анализа процесса разделения жидких фаз. Первые инструменты (PLAAC, LARK, R+Y, CatGranule) для предсказания таких белков основаны на использовании одного-двух признаков, например, на основании наличия prion-like доменов, пропорций аргининов и тиразинов, похожести белков на белок DDX4 и др. Однако начиная с 2020-ых начали появляться алгоритмы на основе машинного обучения, например, deePhase, PSAP и др. учитывающие ряд физико-химических свойств белков.

Целью данной работы является анализ белков человека, участвующих в разделение жидких фаз, построение классификатора машинного обучения и его применение для анализа гибридных онкобелков. Для обучения классификатора использовались табличные данные с физико-химическими свойствами белков, фракциями неупорядоченных регионов и регионов низкой сложности (100 признаков). В качестве положительного класса были взяты каплеобразующие при физиологических условиях и концентрациях белки из литературных данных. Негативный класс - белки человека не имеющие неупорядоченных регионов по предсказаниям AlphaFold 2.0 (на основе значений pLDDT). После проведения кластеризации белковых последовательностей с пороговым значением 0.4, размер каждого класса составил 64 белка. Модель классификатора - градиентный бустинг над решающими деревьями (XGBoost). Точность построенного классификатора на тестовых данных составила 0.91 (std 0.05). При анализе модели были выявлены 20 важных признаков. Была рассмотрена 7061 последовательность гибридных онко-белков из базы FusionGDB, из них 5614 (79.5%) классифицируются как каплеобразующие. 70% каплеобразующих белков образованы при участии двух разделяющих фазы компонентов, 29% белков при участии одного каплеобразующего компонента. Интересны 48 последовательностей, обладающих свойствами разделения фаз при том, что исходные компоненты такими свойствами не обладали.

Исследование выполнено за счет гранта Российского научного фонда № 18-74-10006-П, https://rscf.ru/project/18-74-10006/.