Секция «Физико-химическая биология»

Комбинированный квантово-химический и основанный на знаниях подход к прогнозированию проницаемости биологических мембран

Научный руководитель – Пидько Евгений Александрович

Тихонова Анастасия Евгеньевна

Студент (магистр)

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия

E-mail: tikhonova@scamt-itmo.ru

Большинство существующих в настоящее время подходов к прогнозированию проницаемости соединений через клеточную мембрану являются математическими и механистическими моделями, варьирующимися от молекулярных до макроскопических описаний.Наиболее распространенной вычислительной практикой является представление мембран 1-3 липидами. При этом известно, что мембранный состав *in vivo* оказывает большое влияние на внутренние процессы клетки. Например, как показано Bogdanov et al, использование мембран, содержащих только фосфатидилхолин,может привести к потере нативной структуры белка для некоторых трансмембранных белков, поэтому правильное сворачивание в мембрану, как предполагается, сильнозависит от липидного состава [Bogdanov et al, 2010]. Кроме того, патологические состояния также сильно влияют на структуру мембран, в какой некоторые липиды, такие как фосфатидилсерины, рассматриваются в качестве биомаркеров рака [Szlasa et al,2020]. Изменение биофизических свойств мембран за счет появления новых липидов приводит к нарушению сигнального пути и медицинской проблеме высокой химиотерапевтической резистентности. Целью данного исследования является создание предсказательной модели проницаемости мембран для малых органических молекул, учитывающей сложность иразнообразие конкретных типов клеток и тканей и включающей как нормальные физиологические, так и патологические состояния. Новизна исследования заключается создании модели, для обучения которой используются данные коэффициента проницаемости (LogPerm), полученного из профиля свободной энергии и определяемого как Perm = -J /(Cout - Cin), где J - поток соединения через мембрану, а Cout - Cin - градиентконцентрации между двумя компартментами, разделенный мембраной. Данные, включающие в себя типы тканей и клеток и значения коэффициента проницаемости для малых молекул, были получены в базе данных MolMeDB,содержащей информацию о 14879 соединениях. Исходный набор данных был разделенна подкатегории “Проницаемость тканей кожи”, “Проницаемость тканей желудочно-кишечного тракта”, “Проницаемость почечного монослоя MDCK”, “Проницаемость клеток колоректальной карциномы Caco-2”, “Проницаемость тканей желудочно-кишечного тракта”, “Проницаемость тканей глаза”, и общую категорию, включающую в себя данные с неопределенным типом мембраны. Данное разделение позволит рассмотреть как и нормальные физиологические, так и патологические состояния.Экспериментальные измерения были отделены от теоретических расчетов, которые небыли включены в набор данных. Вклад отдельных химических групп в проницаемость был оценен с помощьюбиблиотеки sirms-spci для выявления различий в проницаемости между различными наборами данных. Для построения моделей машинного обучения был рассчитан набор физико-химических дескрипторов с использованием программного обеспечения ACD/ChemSketch и KNIME, а также библиотеки Pybel. Для оценки корреляции между коэффициентом проницаемости и физико-химическими дескрипторами были построеныкорреляционные матрицы, корреляция считалась значимой при p> 0,05. Размерность обучающих данных

была уменьшена с помощью отбора дескрипторов. Выбор дескрипторов для дальнейшего использования в предсказательной модели осуществлялся с помощью итеративных алгоритмов пошаговой регрессии (обратное ипрямое исключение признаков), а также с помощью генетического алгоритма и методом случайного выбора для оценки эффекта метода сортировки. Все расчеты были выполнены в программном обеспечении Knime с использованием узлов LinearRegression Learner и Predictor. Для выбора оптимального алгоритма машинного обучения был использован автоматизированный фреймворк Autogluon, включающий в себя методы Random Forest, Extra Tree, KNN, Light Gradient Boosting Machine, CatBoost model(Gradient Boosting on Decision Trees), Tabular Neural Net Model и Weighted Ensemble (мета-модель, реализующая выбор ансамбля). Вычислительная точность полученных моделей сравнивалась с QSAR-моделями QSAR-HB, QSAR-PSA, QSAR-Volsurf и механистической моделью PerMM.В ходе характеристики вкладов химических групп в проницаемость для всех наборов данных было выявлено отрицательное влияние гидрофильных групп, при этом были обнаружены вариации между различными группами. Так, набор данных“Проницаемость тканей глаза” показывает положительное влияние сульфонамидныхгрупп (которые являются гидрофобными), показывая, что проницаемость между различными типами тканей может значительно различаться. Положительное влияние гидрофобных групп (по фенильной группе) и отрицательное влияние гидрофильных групп также было показано в наборе данных клеток Caco-2.После оценки корреляции между коэффициентом проницаемости и физико-химическими дескрипторами были выделены три основные группы свойств: атомные дескрипторы (количество элементов/тяжелых атомов, молекулярный вес), дескрипторы связей), количество ароматических/вращательных/двойных/тройных связей, общее количество связей) и дескрипторы, связанные с липофильностью (площадь полярной поверхности, количество доноров/акцепторов водородных связей, поляризуемость связей/атомов). Важность разделения наборов данных на специфические группы тканей и клеток была продемонстрирована путем оценки предсказательных характеристикполученных моделей. Набор данных для неспецифического типа мембраны показал высокие значения ошибок RMSE (начиная с 0.913 и до 1.159). Напротив,кластеризованные наборы данных показали относительно низкие значения ошибок(0.402 для проницаемости тканей кожи, 0.463 для клеток Caco-2 и 0.366 для проницаемости тканей глаза). Не было выявлено значимых различий разницы междупредсказательной способностью методов машинного обучения, т.к. большинство моделей имели схожие значения метрики RMSE. Это может указывать на то, что предсказательная способность модели в большей степени зависит от качества набораданных, нежели от сложности алгоритма. В сравнении с методами QSAR, нацеленными на расчет проницаемости мембран Caco-2, разница между экспериментальными и расчетными значениями была выше, чем в предложенном алгоритме машинного обучения, в то время как модель PerMM показала лучшую предсказательную способность по отношению к неспецифическому типу мембран.Выводы:Высокая предсказательная способность алгоритма по отношению ккластеризованным наборам данных говорит о необходимости использования тканеспецифических моделей для более точного расчета коэффициента проницаемости иуглубленного изучения различий между липидным составом и физико-химическими свойствами различных типов биологических мембран. Данный комбинированный подход, основанный на машинном обучении квантово-химических данных, может быть использован для дальнейшего изучения в вычислительной химии.

Источники и литература

1) Bogdanov, M. et al. Plasticity of lipid–protein interactions in the function and topogenesisof the membrane protein lactose permease from Escherichia coli // Proc. Natl.

Acad. Sci.,2010. No 107 (34). p. 15057-15062

2) Szlasa, W. et al. Lipid composition of the cancer cell membrane. // J Bioenerg Biomembr,2020. . No 52, p. 321–342