



МОДЕЛИРОВАНИЕ ВЗАИМОДЕЙСТВИЯ БЕЛОК-БЕЛОК С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ: ОТ ПРЕДСКАЗАНИЯ ПАР ДО ДИЗАЙНА ТЕРАПЕВТИЧЕСКИХ АГЕНТОВ

Волков Андрей Игоревич

Доцент кафедры сравнительной грамматики, Московский государственный университет
г. Москва Россия

Аннотация

Данная статья посвящена глубокому анализу применения методов машинного обучения (МО) и искусственного интеллекта (ИИ) для моделирования и предсказания взаимодействий белок-белок (ВББ), которые являются центральными процессами в молекулярной биологии и патогенезе заболеваний. ВББ формируют сложную сеть, лежащую в основе клеточного сигналинга, метаболизма и иммунного ответа. Традиционные экспериментальные методы (дрожжевая двухгибридная система, масс-спектрометрия) являются трудоемкими и дорогостоящими. Рассматривается, как МО, включая опорные векторные машины (SVM), случайные леса и глубокое обучение (нейронные сети), трансформирует эту область, используя для предсказания обширные геномные, структурные и эволюционные данные. Особое внимание уделяется применению глубоких сверточных и графовых нейронных сетей для предсказания структуры комплексов (докинга) и определения специфических сайтов связывания (эпитопов). Анализируются ключевые вызовы, связанные с несбалансированностью данных (большинство потенциальных пар не взаимодействует) и необходимостью интеграции гетерогенных признаков. Сделан вывод о том, что МО не только повышает скорость и точность предсказания ВББ, но и открывает путь к рациональному дизайну лекарственных препаратов, нацеленных на модуляцию патологических взаимодействий.

Ключевые слова: Взаимодействие белок-белок (ВББ), Машинное обучение, Глубокое обучение, Секвенирование, Докинг, Предсказание взаимодействия, Графовые нейронные сети, Персонализированная медицина, Дизайн лекарств, Протеомика.

Введение

Жизнедеятельность любой клетки определяется работой тысяч белков, которые не функционируют изолированно, а образуют динамические, многокомпонентные комплексы, формируя сложную иерархическую сеть взаимодействий.

Взаимодействия белок-белок (ВББ) являются краеугольным камнем всех без исключения клеточных процессов — от репликации ДНК и передачи сигналов через мембрану до мышечного сокращения и апоптоза. Нарушение ВББ лежит в основе патогенеза большинства человеческих заболеваний, включая рак, нейродегенеративные расстройства и вирусные инфекции. Понимание и картографирование всей сети ВББ (протеома) — это фундаментальная задача постгеномной эры.

Традиционные подходы к идентификации ВББ (например, методы на основе дрожжевой двухгибридной системы, соосаждение, масс-спектрометрия) обладают высокой точностью, но являются чрезвычайно трудоемкими, дорогими и не позволяют быстро проанализировать весь протеом организма. Это создало огромный разрыв между количеством известных белковых последовательностей и числом экспериментально подтвержденных взаимодействий. В последние годы методы машинного обучения (МО) и глубокого обучения (Deep Learning) стали ключевым инструментом для преодоления этого разрыва. Используя огромные объемы накопленных геномных, транскриптомных и структурных данных, МО способно предсказывать тысячи новых ВББ с высокой точностью, а также моделировать пространственную структуру комплексов. Цель данной работы — систематизировать основные подходы МО к моделированию ВББ, рассмотреть их применение в предсказании структуры и дизайна лекарств, а также проанализировать основные вызовы, стоящие перед этой быстро развивающейся областью.

Подходы Машинного Обучения к Предсказанию ВББ

В зависимости от задачи — предсказание факта взаимодействия или предсказание структуры комплекса — используются различные классы алгоритмов и наборы входных данных (признаков).

Предсказание Бинарного Взаимодействия (Да/Нет): Самая простая и распространенная задача — классификация потенциальной пары белков как "взаимодействующей" или "невзаимодействующей". Для этого используются традиционные алгоритмы МО: опорные векторные машины (SVM), случайные леса (Random Forest) и градиентный бустинг. Входные данные для таких моделей являются гетерогенными и включают:

Последовательностные признаки: Аминокислотный состав, частота встречаемости ди- и трипептидов, длина последовательности и наличие сигнальных пептидов. Эти признаки легко извлекаются из общедоступных баз данных.

Эволюционные признаки: Совместная эволюция (co-evolution) белков. Если два белка эволюционируют синхронно (изменения в последовательности одного компенсируются изменениями в другом), это часто указывает на функциональную или физическую связь.

Структурные признаки: Если известны 3D-структуры белков, используются такие метрики, как площадь поверхности, гидрофобность, распределение зарядов.

Функциональные признаки: Сходство по функции, совпадение по клеточной локализации или участие в одних и тех же метаболических путях.

Объединение этих гетерогенных признаков (fusion) позволяет МО-моделям достигать высокой точности (AUC до 90%) в предсказании факта взаимодействия.

Предсказание Взаимодействий на Основе Глубокого Обучения: Глубокое обучение, особенно сверточные нейронные сети (CNN) и графовые нейронные сети (GNN), показало революционные результаты, поскольку оно способно автоматически извлекать сложные, нелинейные признаки из сырых данных (последовательностей или структур) без необходимости ручного инжиниринга признаков.

Сверточные сети (CNN): Используются для анализа последовательностей. Они могут автоматически определять важные мотивы или паттерны в аминокислотных последовательностях, которые служат сайтами связывания, и сравнивать эти паттерны между двумя белками для предсказания их физического контакта.

Графовые нейронные сети (GNN): Это наиболее мощный инструмент для анализа ВББ. Белки представляются как узлы в графе, а их взаимодействие — как ребро. GNN могут обрабатывать иерархическую структуру протеомных сетей, предсказывая не только факт взаимодействия между двумя узлами, но и анализируя влияние этого взаимодействия на всю клеточную сеть, а также предсказывая новые взаимодействия на основе топологии графа.

Моделирование Структуры Комплексов (Молекулярный Докинг)

Самой сложной задачей в структурной биоинформатике является предсказание трехмерной структуры комплекса, образуемого двумя взаимодействующими белками, на основе их индивидуальных структур — процесс, известный как молекулярный докинг.

Традиционный Докинг и Его Ограничения: Классические алгоритмы докинга (например, ClusPro, ZDOCK) используют алгоритмы перебора, чтобы найти наиболее энергетически выгодное взаимное расположение белков. Эти методы, однако, очень чувствительны к изменениям конформации (гибкости) белков при связывании и требуют колоссальных вычислительных мощностей для анализа всех возможных ориентаций.

Докинг, Управляемый Машинным Обучением: МО-подходы значительно повышают скорость и точность докинга, решая две основные задачи:

Ранжирование Поз: После генерации тысяч возможных поз связывания традиционными алгоритмами, МО-модели (часто на основе бустинга или

глубоких сетей) обучаются ранжировать эти позы, выбирая ту, которая наиболее соответствует нативной структуре. В качестве признаков используются геометрические, электростатические и гидрофобные характеристики контакта.

Предсказание Сайтов Связывания (Хотспотов): МО может предсказать специфические "горячие точки" (hotspots) — остатки на поверхности белка, которые с наибольшей вероятностью будут участвовать в формировании контакта. Для этого используются сверточные нейронные сети, которые обрабатывают 3D-данные о поверхности белка как изображения. Это значительно сужает область поиска для традиционных алгоритмов докинга, сокращая время вычислений с дней до часов и повышая точность определения нативной конформации комплекса.

Вызовы и Перспективы в Дизайне Лекарств

Применение МО для моделирования ВББ имеет огромный потенциал в разработке лекарственных средств, но сталкивается с рядом технологических вызовов.

Вызовы Несбалансированности и Гетерогенности Данных: Главная проблема в обучении моделей предсказания ВББ — это **несбалансированность классов**. Количество взаимодействующих пар белков (положительный класс) составляет менее 1% от общего числа возможных пар в протеоме. Недостаток надежных отрицательных примеров (пар, которые точно не взаимодействуют) может привести к смещению модели. Кроме того, требуется интеграция чрезвычайно гетерогенных данных (структурных, геномных, транскриптомных), что требует сложных архитектур глубокого обучения, способных обрабатывать разные типы входных данных одновременно.

Дизайн Модуляторов Взаимодействий: Ключевая перспектива — это рациональный **дизайн лекарств**, которые могут либо **ингибировать** (блокировать) патологическое ВББ (например, связывание вирусного белка с клеточным рецептором), либо **активировать** (усиливать) полезное взаимодействие. МО позволяет не просто предсказывать, а **проектировать** пептидные или малые молекулы, которые будут комплементарны сайтам связывания на белковых поверхностях. Глубокие генеративные модели (например, вариационные автокодировщики, VAEs) обучаются генерировать новые молекулярные структуры *de novo*, оптимизированные для связывания с целевым интерфейсом белка. Это открывает путь к созданию совершенно новых классов терапевтических средств — модуляторов протеомных сетей.

Заключение

Методы машинного обучения и глубокого обучения совершили качественный скачок в моделировании взаимодействий белок-белок, переведя эту область из трудоемкой экспериментальной науки в высокопроизводительную вычислительную дисциплину.

Использование графовых нейронных сетей для анализа протеомных сетей и глубоких сверточных сетей для высокоточного докинга позволяет не только быстро картировать ВББ, но и решать сложные структурные задачи. Несмотря на сохраняющиеся проблемы, такие как несбалансированность данных и необходимость валидации предсказаний, МО уже становится ключевым инструментом в рациональном поиске новых лекарств и дизайне терапевтических агентов. Интеграция этих вычислительных методов в клинические исследования ускорит понимание молекулярных основ заболеваний и, в конечном итоге, будет способствовать развитию персонализированной медицины.

Литература

1. Lensink, M. F. *Blind docking of proteins to proteins: New insights from CAPRI*. Proteins: Structure, Function, and Bioinformatics, 2017.
2. Skerman, A., Li, W., Kulkarni, V. *Machine learning in protein-protein interaction prediction*. Briefings in Bioinformatics, 2016.
3. Zeng, H., Yang, H., Wang, J. *Deep learning for protein-protein interactions*. Current Opinion in Structural Biology, 2021.
4. Ragoza, M., Hochuli, J., Idrobo, E. *Filter-based deep convolutional neural networks for predicting protein-ligand binding affinity*. Chemical Science, 2017.
5. Gligorijević, V., Stokach, A., Šegota, E. *Deep learning for protein-protein interaction prediction*. Bioinformatics, 2021.
6. Смирнов, О. В. *Биоинформатика: Основы и методы*. Наука, 2023.
7. Киселев, И. А. *Машинное обучение в структурной биологии*. Информационные технологии, 2024.