УДК 004.85

ПРЕДСКАЗАНИЕ ТРЕТИЧНОЙ СТРУКТУРЫ БЕЛКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Кочкова С.В., Университет ИТМО, Санкт-Петербург

Научный руководитель – кандидат технических наук, доцент факультета информационных технологий и программирования Сергушичев А.А., Университет ИТМО, Санкт-Петербург

В рамках данной работы были изучены существующие методы предсказания трехмерной структуры белка методами машинного обучения, где объектом предсказания являлась матрица попарных расстояний между каркасными атомами. Также были изучены свойства такой матрицы и доказано, что можно предсказывать уменьшенную версию матрицы, что должно улучшить точность предсказания. Была собрана база антител и построена модель машинного обучения.

Введение. На данный момент задача построения трехмерной структуры белка остается актуальной, особенно это касается структуры антител. Существует множество подходов к решению этой задачи, но ни по одному из этих подходов пока не получилось получить результатов, достаточно близких к действительной структуре антитела. Антитело состоит из трех вариабельных регионов и четырех константных, и если константные варьируются слабо и их структуру предсказать легче, то для петель, чья вариабельность очень высока, предсказать структуру сложнее. При этом для первой и второй петель в современном мире уже написано программное обеспечение, позволяющее достаточно точно их предсказывать, но третья петля вариативна настолько, что успеха в её предсказании не достигли. Также для произвольных белков тоже нет оптимального решения по предсказанию структуры. В рамках данной работы планировалось изучить свойства матрицы попарных расстояний, которую можно использовать в качестве объекта предсказания методами машинного обучения, чтобы более точно предсказывать её, а после собрать базу данных и написать модель для предсказания.

Основная часть. Трехмерная структура белка определяется координатами атомов, из которых состоят аминокислоты. При этом основной каркас любой аминокислоты состоит из атомов {N, CA, C}, и если предсказать их координаты, остальные атомы восстанавливаются с помощью существующих подходов. Поэтому если предсказать атом CA, остальные потом можно восстановить. При этом идея изученного подхода к предсказанию этих атомов с помощью модели машинного обучения заключается в том, чтобы рассматривать последовательности аминокислот как большое количество попарных расстояний между этими атомами. Для использования этой идеи сначала было решено проверить разные возможные свойства матрицы попарных расстояний.

Как оказалось, матрица квадратов попарных расстояний имеет ранг не больше пяти в трехмерном пространстве, в котором существует структура белка. На самом деле ранг матрицы равен минимуму из количества точек и пяти почти всегда. Что значит “почти всегда” — основываясь на существующей статье, в трехмерном пространстве ранг матрицы будет равен 4, а не 5, только если все точки лежат на плоскости, но не на линии или окружности, что в представленном случае предсказания структуры белка не будет выполняться никогда. Таким образом, матрица попарных расстояний точно может быть определена матрицей размера 5 x n, и методами машинного обучения необходимо будет предсказать всего O(n) чисел, а не O(n^2), как это делается в оригинальной статье. Следующий вопрос – каким образом необходимо строить матрицу так, чтобы из вида 5 x n можно было восстановить первоначальную матрицу расстояний.

Изучив все свойства собственных чисел и векторов матрицы попарных расстояний, было доказано, что её можно представить в уменьшенном виде, где каждая из пяти строчек будет равна произведению корня из модуля собственного числа и соответствующего собственного

вектора. Также известно, что все собственные числа матрицы попарных расстояний отрицательные, за исключением одного – самого большого по модулю. Таким образом, при перемножении этого уменьшенного вида матрицы на её транспонированный вид и при умножении на знаки соответствующих собственных чисел, полная матрица попарных расстояний может быть точно восстановлена.
Учитывая сделанные выводы, теперь можно предсказывать не целую матрицу попарных расстояний, а её уменьшенную версию, потому что было доказано, что из уменьшенной версии изначальную можно однозначно восстановить. Далее была собрана база антител, и на её основе планируется реализовать модель машинного обучения для предсказания петель. Пока что результатов не было получено, но в данный момент ведутся работы над построением модели. Также были продуманы метрики для оценки точности модели – среднеквадратичное отклонение от оригинальной структуры петель, такое же отклонение для структур каркасных регионов и для целой структуры.

Выводы. В случае достижения достаточно высокой точности в предсказании, реализованная модель будет использоваться для предсказания структуры антител в компании ЗАО «Биокад», что поможет более точно вычислять физико-химические свойства и функции антител. А далее можно будет использовать эту же модель для предсказания структуры любого белка, потому что она никак не заточена под формат антител.