Санкт-Петербургский Государственный Университет

Кафедра системного программирования

Группа 18.Б11-мм

Цириков Семен Алексеевич

Улучшение качества эпитопного картирования методами глубокого обучения

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.ф.-м.н. Граничин О.Н.

Рецензент:
к.ф-м.н. Ерофеева В.А.

Санкт-Петербург
2022

SAINT PETERSBURG STATE UNIVERSITY

Software engineering

Simon Tsirikov

Improving Quality Of Epitope Mapping By Deep Learning Methods

Bachelor's Thesis

Scientific supervisor:
Dr. of Sci. Oleg Granichin

Reviewer:
C. of Sci. Victoria Erofeeva

Saint Petersburg
2022

Оглавление

Введение

Множество процессов в организме человека сопряжено со взаимодействием белков. Одним из таких процессов является ответ иммунной системы на проникновение потенциально опасных агентов в организм, которое сопровождается связыванием антигена и антитела. Антигеном в классическом смысле называют любое вещество, которое организм классифицирует как чужеродное или потенциально вредоносное. Антителом же называют молекулу, производимую иммунной системой или искусственным путём для уничтожения антигена.

При разработке лекарственных средств одним из этапов исследования является компьютерное моделирование молекул с целью выявить их способность к связыванию и образованию комплексов, называемому докингом. Подзадачей докинга является эпитопное картирование – определение той части молекулы антигена, к которой антитело будет присоединяться.

Точность существующих подходов к решению задачи эпитопного картирования в настоящий момент ограничена [13]. Для её повышения применяются различные методы машинного обучения, в частности глубокое обучение, использующее искусственные нейронные сети [10]. Модели, предложенные до настоящего времени, показывают высокое значение метрики recall[1], но низкое значение precision. Эта ситуация релевантна прикладным задачам, так как увеличивает скорость выполнения задачи и снижает стоимость разработки лекарств – меньшее число предсказаний нужно перепроверять в лабораториях.

Для облегчения доступа исследователей к результатам своей работы, широко используемым подходом является публикация прототипа системы с интерфейсом, через который можно дать на вход свои данные к предобученной модели и получить результаты. Такой подход оказывается существенно более удобным для применения в отличие от идеи опубликовать исходный код без обучающей выборки, что создаёт дополнительные трудности при воспроизведении полученных результатов.

[1] $recall = \frac{TP}{TP+FP}, precision = \frac{TP}{TP+FN}$, где TP, FP и FN – количество истинно положительных, ложноположительных и ложноотрицательных экспериментов соответственно

1. Постановка цели и задач работы

Целью работы является создание модели машинного обучения для решения задачи эпитопного картирования с метрикой качества precision, превосходящей аналоги.

Для достижения указанной цели были поставлены следующие задачи:

- Сделать обзор существующих моделей.
- Выбрать набор данных для обучения.
- Спроектировать и реализовать модель глубокого обучения.
- Провести эксперименты, сравнить результаты с другими моделями.
- Разработать прототип для использования предобученной модели.

2. Обзор литературы и методов

Из-за вычислительной сложности, вызванной большим размером белковых молекул, на практике решить задачу полным перебором различных конформаций не представляется возможным, а в условиях отсутствия информации об антигене такой подход не применим. Методы, использовавшиеся для приближенного решения задачи до применения машинного обучения, описаны в [13].

Для того, чтобы выбрать подходящую модель обучения, произведён обзор нескольких недавних работ, показавших наилучшие результаты по метрикам качества и предложивших перспективные методы для кодирования информации о молекулах и обучении на этих данных.

2.1. PECAN – графовые сверточные сети для комплексов

Авторы PECAN (Paratope and Epitope prediction with graph Convolution Attention Network) [12] предлагают использовать графовую модель представления молекул вместо пространственной, так как считают, что связи между атомами можно охарактеризовать рёбрами, что лучше передаёт локальные свойства регионов, чем перебор по координатам. Они используют только структурную информацию, строя граф таким образом, что рёбра есть между любыми двумя атомами, расстояние между которыми менее 10Å[2].

Примечательно использование техники Transfer Learning [6], которая позволяет обучиться на большем множестве белковых комплексов, чем только антиген-антитело (исходя из предположения, что сходные участки белковых молекул проявляют схожие свойства в разных белках), а затем применить результат обучения к нужной задаче.

Однако, как отмечается в исследовании проекта ATOM3D [3] (коллекция данных о трёхмерной структуре биомолекул), для представления свойств больших молекул, какими являются белки, графы являются

[2]Å (ангстрем) – единица измерения длины, равная 10^{-10} метра

неэффективным средством.

2.2. PInet – геометрические сверточные сети для комплексов

В PInet (Protein Interface Network) [7] используется геометрическое представление атомов, а в качестве характеристических свойств используются коэффициенты гидрофобности и электростатики для каждого атома. Так же, как и PECAN, эта модель делает предсказание расположения эпитопа для пары антитело-антиген. В [7] также было предложено использовать независимые контрольные наборы данных для валидации обученной модели, такие как DBD5 [15] (docking benchmark databasse version 5) и PRISM [11] (Protein Interactions by Structural Matching). Для разметки эпитопа в обучающих данных используется общепринятое значение расстояния в 5Å.

2.3. Epitope3D – ансамбль для антигенов

Модель Epitope3D, предложенная в [17], обучена с использованием самого большого открытого набора данных о белках PDB [18] (Protein Data Bank), алгоритм обучения – AdaBoost [9] (Adaptive Boosting). Авторы предложили большое количество свойств, выбирая из них дающие наибольший вклад в коэффициенты с помощью методики жадной выборки свойств [16]. После обучения модели, на вход для предсказания требуется только антиген.

2.4. Molformer – трансформер для представления молекул

В [1] авторы описали модель, созданную не для непосредственного решения задачи эпитопного картирования, а для кодирования информации о молекуле. Авторы считают, что предложенный ими подход снижения размерности модели может помочь в увеличении качества и снижении

вычислительной сложности для задач предсказания свойств молекул, о чём свидетельствуют их эксперименты. Алгоритм основан на архитектуре Transformer, предложенной в [4], и позволяет создавать модели молекул, в которых большое количество атомов будет заменено на вектор вещественных чисел выбранной длины, представляющий свёртку некоторой области вместе с её локальными свойствами.

3. Модель эпитопного картирования

Основываясь на результатах предыдущих исследований, спроектирована архитектура собственной модели, модель реализована, выбраны тренировочные данные и произведено обучение, сделаны выводы о возможностях модели.

3.1. Выбор данных для обучения

Помимо наборов данных, упомянутых ранее, рассматривался репозиторий Anbase [2] , спроектированный для задачи обучения алгоритмов, предсказывающий свойства взаимодействия антитела и антигена. По результатам сравнения, приведённом в Таб. 1, Anbase был выбран для использования при обучении собственной модели.

База данных	Размер	Тип
PRISM	6001	Белки
PDB	10714	Белки
DBD5	409	Антигены
Anbase	570	Антигены

Таблица 1: Сравнение баз данных, используемых в задачах взаимодействия антитела и антигена

В Таб. 1 под размером понимается количество связанных структур белок-белок, представленных в базе данных, в поле тип написано “белки” в случае произвольных белков, “антигены”, если содержатся только структуры, для которых известно, что они распознаются организмом как антигены. Было принято решение не использовать методику Transfer Learning и обучаться только на антигенах, так как есть основания полагать, что антигены устроены более сложно, чем белки в среднем, и у них не очень много общих шаблонов в структуре.

3.2. Описание разработанной модели

Предложенная модель действует следующим образом: сперва данные извлекаются из файла в формате PDB, откуда извлекается информация

о всех атомах, входящих в антиген, и как атомы сгруппированы в аминокислоты.

Размер аминокислот недостаточно большой, чтобы охарактеризовать некоторый регион, а также возникает проблема неоднородности, разные аминокислоты могут иметь различное количество атомов, в то время как на вход алгоритмов ожидаются данные одинакового размера. Чтобы решить эту проблему, строится структура данных k-d tree [5], которое помогает быстро находить близко расположенные атомы, с помощью которой антиген разбивается на регионы. При обучении каждому такому региону будет присвоен один из классов: 0 для тех регионов, которые не входят в эпитоп, 1 – входят.

На данном этапе регион не закодирован и содержит ту же информацию, что в PDB файле. Далее он подаётся на вход трансформеру, описанного в Molformer [6], чья задача построить свёртку этого региона в более компактное представление, характеризующее некоторые структурные свойства молекулы.

На выходе из трансформера получается вектор действительных чисел фиксированного размера, с которым работают следующие модули модели, состоящие из двух линейных слоёв и функции активации. Результатом их работы является пара чисел, которые соответствуют вероятностям отнесения к классу 0 или 1 соответственно, предсказанный класс определяется по максимальной вероятности.

Для обучения модели использовались в качестве функции потерь перекрёстная энтропия и оптимизатор, основанный на методе ADAM [14] (Adaptive Moment Estimation), для тренировочной выборки было использовано 70% данных, 10% – для валидации, 20% – для тестирования.

3.3. Анализ результатов

Полученные результаты, приведённые в Секции 4, свидетельствуют о том, что использованной информации недостаточно для решения задачи эпитопного картирования. Для получения качественно более высоких значений метрик необходимо использовать более сложную модель. Ско-

рость обучения относительно медленная, функция потерь убывает, как показано на Рис.1 (обучение остановлено по принципу ранней остановки, когда значение функции потерь начинает возрастать, для борьбы с переобучением), при наличии больших вычислительных мощностей можно увеличить размер окружения.

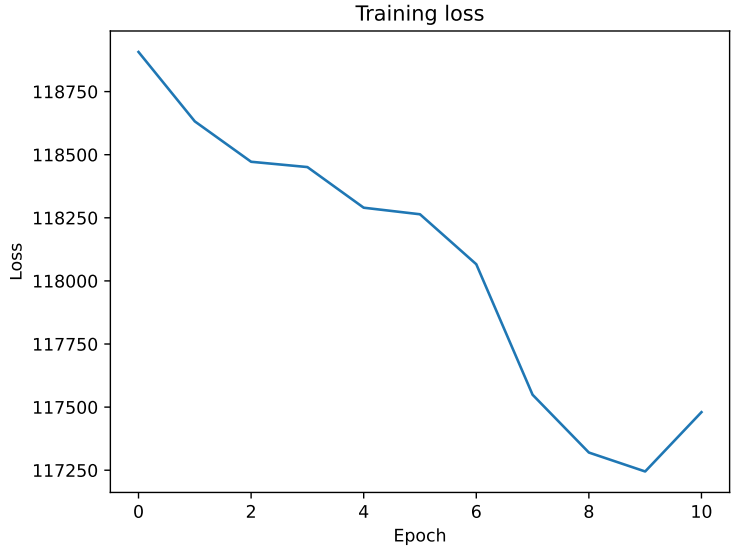

Рис. 1: График функции потерь предложенной модели по эпохам

4. Прототип системы

Для упрощения использования разработанной системы опубликована предобученная модель и описан интерфейс взаимодействия с программой, а так же проведено сравнение качества предсказаний с другой моделью.

4.1. Описание интерфейса

На вход программа получает имя файла в формате PDB, в котором хранится информации об антигене. Если файл существует и его содержание соответствует формату, то на выходе программа напечатает в консоль список номеров тех аминокислот, которые по результатам работы модели считаются входящими в эпитоп. С кодом можно ознакомиться в репозитории [8].

4.2. Оценка и сравнение моделей

Так как используется одинаковый формат входа, возможно сделать сравнение с моделью Epitope3D. Были отобраны данные, которые не использовались в обучении ни одной из моделей, и был написаны скрипты для запуска тестов для Epitope3D, которые также представлены в репозитории [8].

При валидации были получены результаты, метрики которых представлены в Таб.2.

Модель	Precision	Recall
Epitope3D	0.084	0.191
Модель	0.153	0.090

Таблица 2: Сравнение метрик качества разработанной модели и Epitope3D

5. Заключение

В ходе работы достигнуты следующие результаты:

- Проведён обзор существующих моделей: PECAN, PInet, Epitope3D, Molformer.
- Выбрана для обучения база данных Anbase.
- Спроектирована и реализована модель глубокого обучения по архитектуре Transformer.
- Проведены эксперименты, сделано сравнение метрик precision и recall с моделью Epitope3D, достигнуто увеличение precision на 80%.
- Для использования предобученной модели разработан прототип системы на языке программирования python.

Список литературы

[1] Wu Fang, Zhang Qiang, Radev Dragomir, Cui Jiyu, Zhang Wen, Xing Huabin, Zhang Ningyu, and Chen Huajun. 3D-Transformer: Molecular Representation with Transformer in 3D Space. — arXiv preprint arXiv:2110.01191, 2021.

[2] Anbase. — 2022. — Access mode: `https://github.com/biocad/anbase`.

[3] Townshend Raphael JL, Vögele Martin, Suriana Patricia, Derry Alexander, Powers Alexander, Laloudakis Yianni, Balachandar Sidhika, Jing Bowen, Anderson Brandon, Eismann Stephan, et al. Atom3d: Tasks on molecules in three dimensions. — arXiv preprint arXiv:2012.04035, 2020.

[4] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need // Advances In Neural Information Processing Systems. — 2017. — Vol. 30.

[5] Bentley Jon Louis. Multidimensional binary search trees used for associative searching // Communications of the ACM. — 1975. — Vol. 18, no. 9. — P. 509–517.

[6] Bozinovski S and Fulgosi A. The influence of pattern similarity and transfer learning upon training of a base perceptron b2 // Proceedings of Symposium Informatica. — 1976. — P. 3–121.

[7] Dai Bowen and Bailey-Kellogg Chris. Protein interaction interface region prediction by geometric deep learning // Bioinformatics. — 2021.

[8] EpitopeMappingPredictor. — 2022. — Access mode: `https://github.com/SimonTsirikov/EpitopeMappingPredictor`.

[9] Freund Yoav and Schapire Robert E. A decision-theoretic generalization

of on-line learning and an application to boosting // Journal of Computer And System Sciences. — 1997. — Vol. 55, no. 1. — P. 119–139.

[10] McCulloch Warren S and Pitts Walter. A logical calculus of the ideas immanent in nervous activity // The Bulletin of Mathematical Biophysics. — 1943. — Vol. 5, no. 4. — P. 115–133.

[11] Baspinar Alper, Cukuroglu Engin, Nussinov Ruth, Keskin Ozlem, and Gursoy Attila. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes // Nucleic Acids Research. — 2014. — Vol. 42, no. W1. — P. W285–W289.

[12] Pittala Srivamshi and Bailey-Kellogg Chris. Learning context-aware structural representations to predict antigen and antibody binding interfaces // Bioinformatics. — 2020. — Vol. 36, no. 13. — P. 3996–4003.

[13] Potocnakova Lenka, Bhide Mangesh, and Pulzova Lucia Borszekova. An introduction to B-cell epitope mapping and in silico epitope prediction // Journal of Immunology Research. — 2016. — Vol. 2016.

[14] Singarimbun Roy Nuary, Nababan Erna Budhiarti, and Sitompul Opim Salim. Adaptive moment estimation to minimize square error in backpropagation algorithm // 2019 International Conference of Computer Science and Information Technology (ICoSNIKOM) / IEEE. — 2019. — P. 1–7.

[15] Vreven Thom, Moal Iain H, Vangone Anna, Pierce Brian G, Kastritis Panagiotis L, Torchala Mieczyslaw, Chaleil Raphael, Jiménez-García Brian, Bates Paul A, Fernandez-Recio Juan, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2 // Journal of Molecular Biology. — 2015. — Vol. 427, no. 19. — P. 3031–3041.

[16] Vafaie Haleh, Imam Ibrahim F, et al. Feature selection methods: genetic algorithms vs. greedy-like search // Proceedings of The International Conference On Fuzzy And Intelligent Control Systems. — 1994. — Vol. 51. — P. 28.

[17] da Silva Bruna Moreira, Myung YooChan, Ascher David B, and Pires Douglas EV. epitope3D: a machine learning method for conformational B-cell epitope prediction // Briefings In Bioinformatics. — 2021.

[18] Berman Helen M, Westbrook John, Feng Zukang, Gilliland Gary, Bhat Talapady N, Weissig Helge, Shindyalov Ilya N, and Bourne Philip E. The protein data bank // Nucleic Acids Research. — 2000. — Vol. 28, no. 1. — P. 235–242.