Уточнение предсказания формы антител с помощью глубокого обучения

Соликов П. Д.

Научный руководитель: к.ф.-м.н. Д. Н. Москвин

Научный консультант: А. А. Шпильман

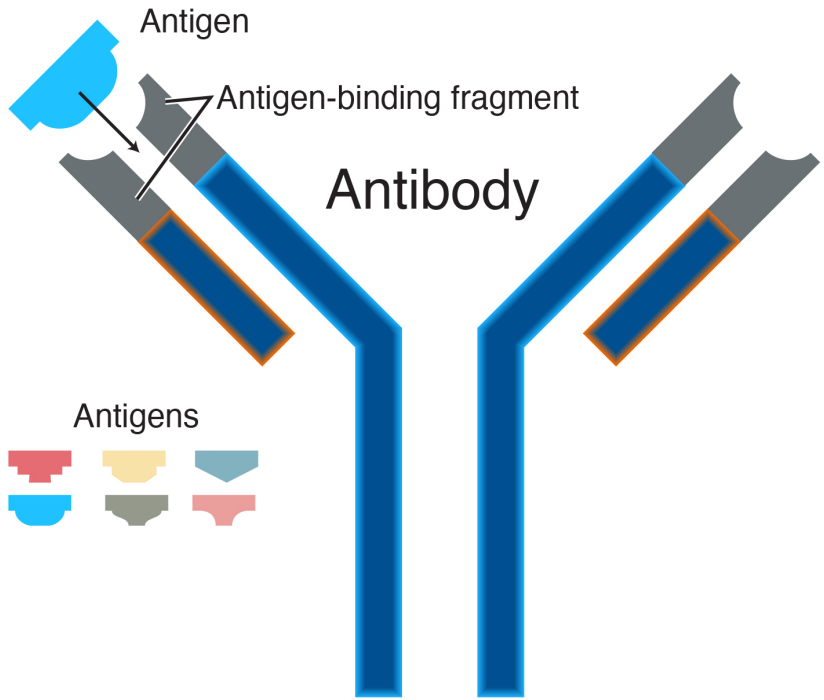
Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург

2020

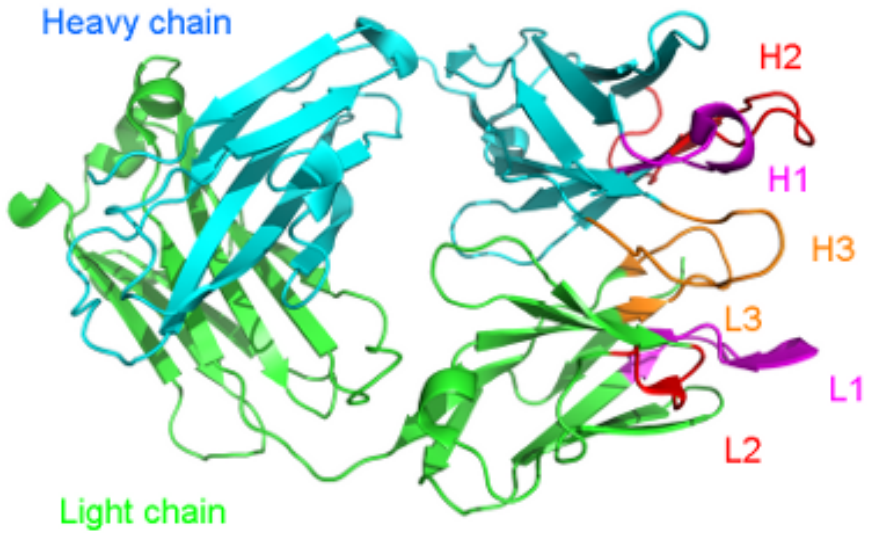
Введение в область

- Антитела – белковые соединения, позволяющие бороться с болезнями
- Искусственно синтезированные антитела широко применяются в фармацевтике
- Синтезирование стоит очень дорого, предварительно нужно отфильтровать только потенциально работоспособные соединения
- Для этого требуется создать алгоритм, способный по белковой последовательности предсказывать форму антитела



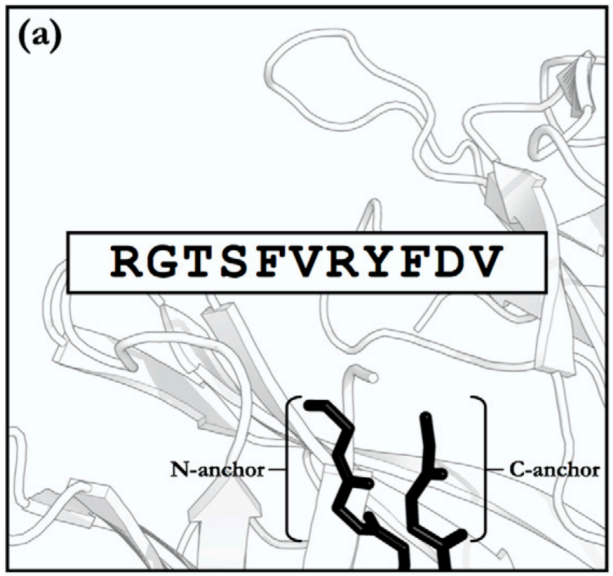
Введение в область

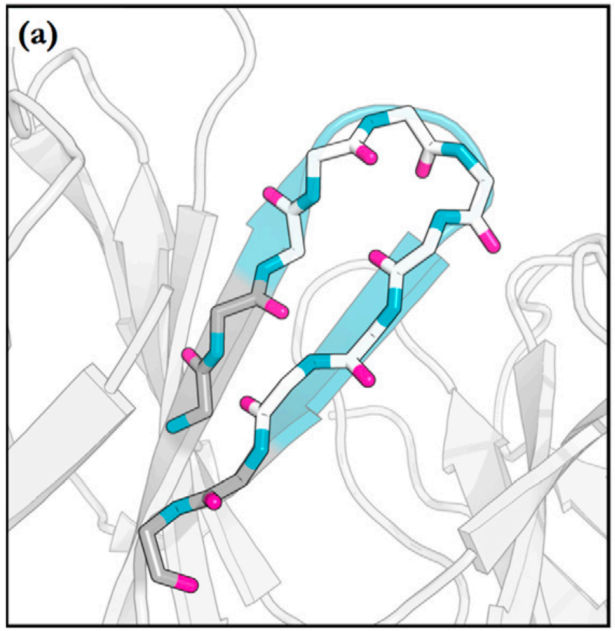
- Большая часть антитела консервативна и ее можно предсказать используя существующие методы
- Плохо моделируются участки, ответственные за связь с антигеном, т.е. самые важные
- Все петли, кроме H3, можно разбить на классы, внутри которых они мало различаются
- Существующие подходы основываются на комбинировании наиболее подходящих известных кусков других антител



Введение в область

- Наиболее вариативный участок антитела CDR H3 – это последовательность от 3 до 25 аминокислотных остатков
- Для каждого аминокислотного остатка нужно узнать координаты 3-х атомов C, C-Alpha и N
- SAbDab – 3935 антител из PDB , 1400 уникальных петель H3





Предсказание формы белков

- CASP – соревнования по предсказанию формы белков, проводятся каждые 2 года, начиная с 1994
- AlphaFold [1], RaptorX [2] – предсказание матрицы попарных расстояний и углов.
- Spider3 [3], Deep3I [4] – предсказание торсионных углов

---

[1] Improved protein structure prediction using potentials from deep learning, Richard Evans, John Jumper, Nature 2020

[2] Distance-based Protein Folding Powered by Deep Learning, Jinbo Xu, PNAS 2019

[3] Single-Sequence-Based Prediction of Protein Secondary Structures and Solvent Accessibility by Deep Whole- Sequence Learning, Rhys Heffernan, Computational Chemistry, 2018

[4] MUFOLD-SS: New Deep Inception-Inside-Inception Networks for Protein Secondary Structure Prediction, Wiley Periodicals, 2018

Существующие решения

- В статье Antibody H3 Structure Prediction[5] приведен обзор всех существующих решений задачи на 2017 год. Все подходы можно разбить на 3 класса: knowledge-based, de novo и гибридные подходы.
  - RosettaAntibody генерирует большое множество возможных петель с помощью случайных вращений исходной структуры. Выбираются структуры, минимизирующие потенциальную энергию.
- Лучшее решение на данный момент описано в статье Jeffrey J. Gray, 2020 [6], улучшающее RosettaAntibody.
- Ни один метод не позволяет достичь точности предсказания, нужной для исследования функций антитела.

[5] Computational and Structural Biotechnology Journal Volume 15, 2017, Pages 222-231

[6] Geometric Potentials from Deep Learning Improve Prediction of CDR H3 Loop Structures, Jeffrey A. Ruffolo, Carlos Guerra , Sai Pooja Mahajan , Jeremias Sulam , Jeffrey J. Gray, 2020 bioRxiv

Цель и задачи

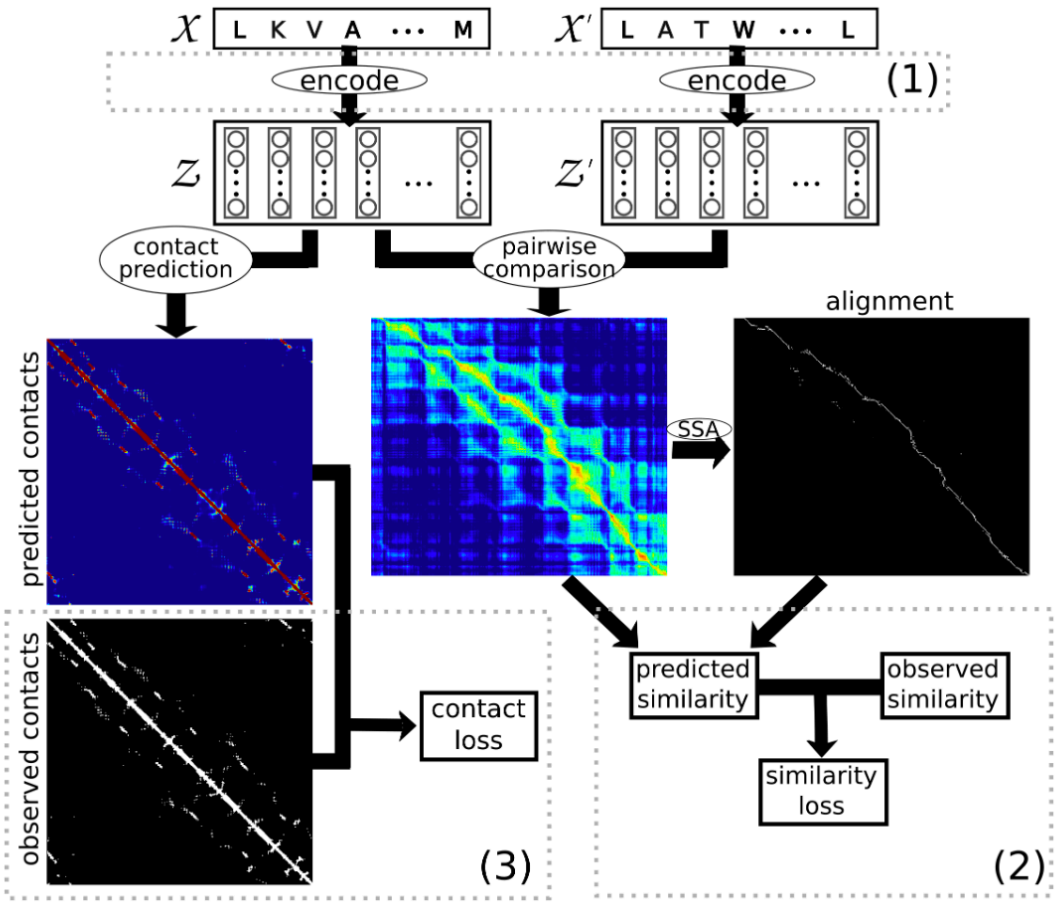
Цель: используя методы глубокого обучения научиться предсказывать трехмерную структуру наиболее вариативного участка антитела – CDR H3

Задачи:

- Выбрать и адаптировать векторное представление для аминокислотной последовательности
- Описать задачу в терминах нейронных сетей
- Провести эксперименты с различными архитектурами глубоких нейронных сетей
- Сравнить полученные результаты с существующими решениями и аналогами

Кодирование последовательности

В статье Bepler [7] векторное представление обучается путем решения задач предсказания контактной матрицы и предсказания попарной схожести белков. Используя полученное векторное представление авторы смогли достичь state-of-the-art результатов в задаче предсказания трансмембранного региона.



[7] Bepler, Tristan and Bonnie Berger. “Learning protein sequence embeddings using information from structure.” ArXiv abs/1902.08661 (2019)

Кодирование последовательности

Статья Heinzinger [8]

- Языковая модель на основе популярного метода ELMO
- 33M неразмеченных последовательностей базы UniProt [9]
- Повторение state-of-the-art результата в задаче предсказания вторичной структуры белка

[8] Heinzinger, Michael & Elnaggar, Ahmed & Wang, Yu & Dallago, Christian & Nachaev, Dmitrii & Matthes, Florian & Rost, Burkhard. (2019). Modeling the Language of Life - Deep Learning Protein Sequences. 10.1101/614313.

[9] https://www.uniprot.org/

Требования к модели

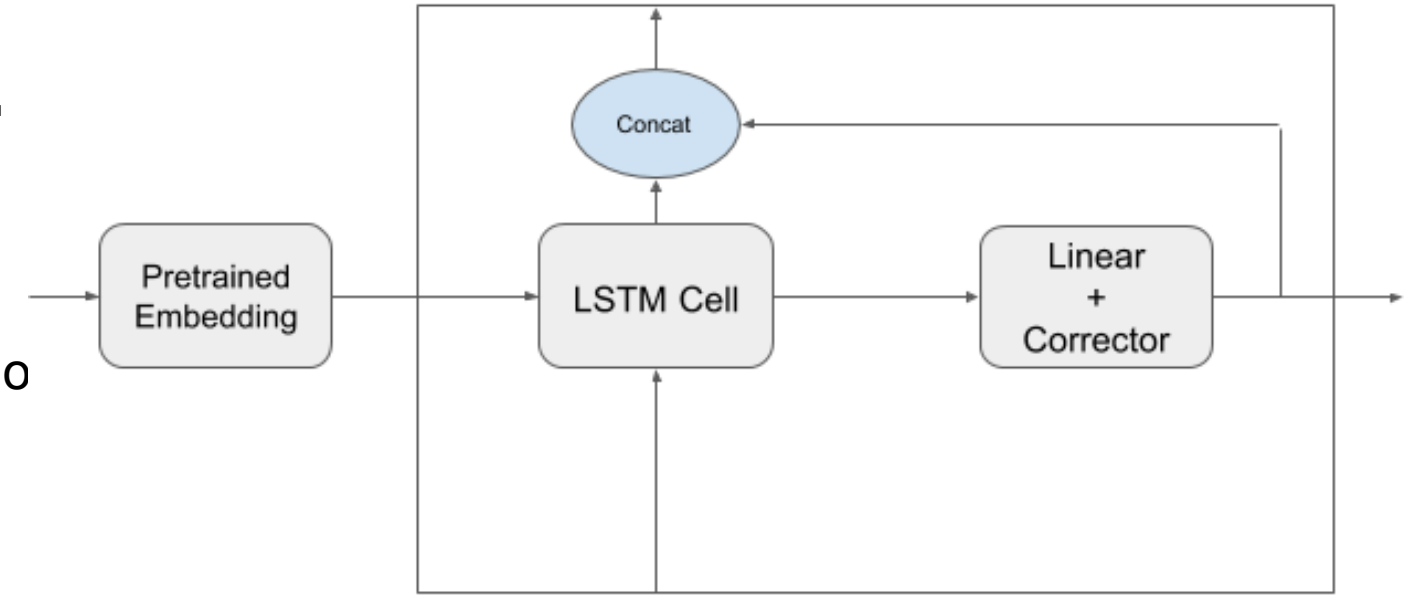
- Модель должна минимизировать отклонение координат атомов предсказанной структуры от реальной (метрика RMSD), при этом сохранив физический смысл
- Минимизация следующей функции ошибок приводит к выполнению требований:
  - MSE координат атомов
  - MSE плоских углов между атомами
  - MSE расстояния между соседними атомами
  - Итоговая функция – сумма

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2$$

$$\mathrm{RMSD}(\mathbf{v},\mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$$

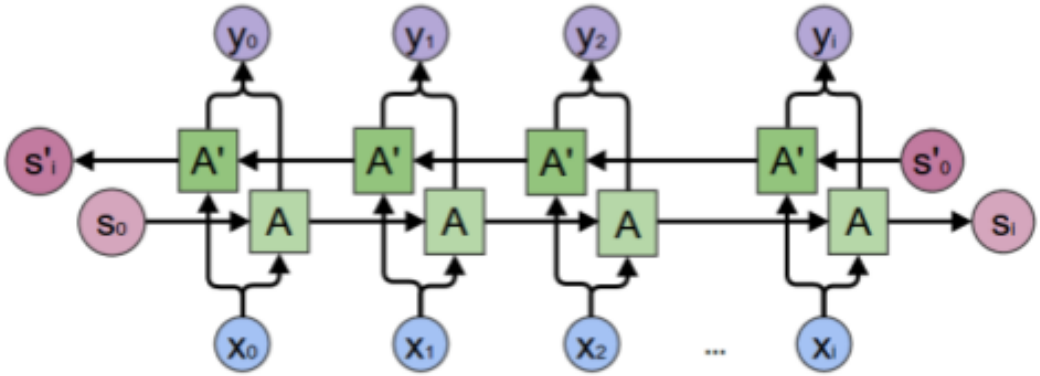
Модель Bidirectional CharLSTM

- Обратный проход без коррекции
- Прямой проход учитывает результат обратно и применяет коррекцию



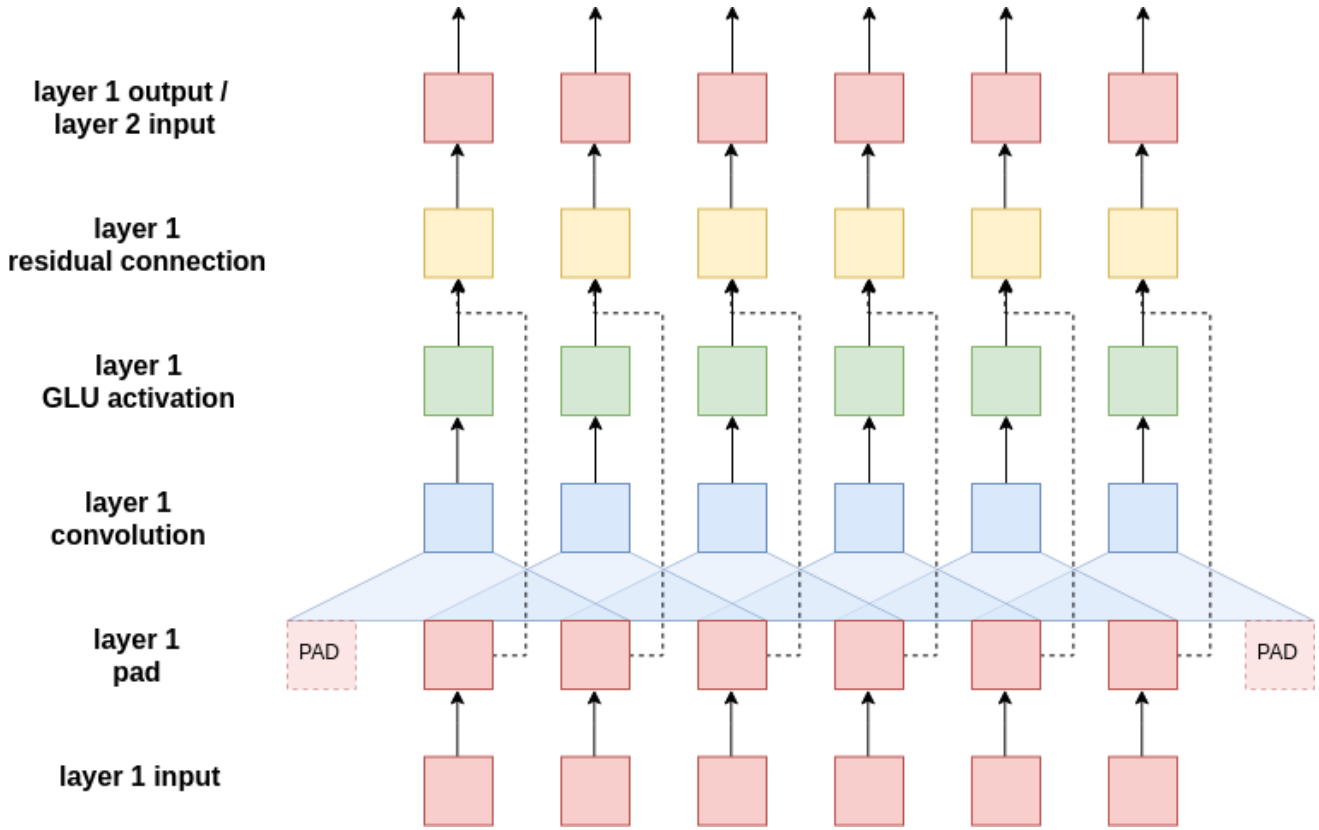
5% SabDab

Embedding	RMSD, Å
Bepler	4.6
Heinzinger	3.2



Модель CNN

- Одномерные сверточные блоки
- Уменьшение размера тензора до 9
- LSTM для коррекции координат



5% SabDab

Embedding	RMSD, Å
Bepler	7.5
Heinzinger	7.7

Тестирование

Соревнование Antibody Modeling Assessment 2[10], в котором участвовали 7 команд, предсказывая 11 заранее не известных структур. Модель CharLSTM с векторным представлением Heinzinger была протестированна на этой выборке.

- Второе место на структурах Ab01, Ab02 и Ab09
- Третье место на структуре Ab010
- Последнее место на структурах Ab08 и Ab011, однако результаты остальных лишь не на много лучше

[10] http://www.3dabmod.com/

Тестирование

Результаты, используя выравнивание

Модель\структура	4kq4	4m43	4m7k	4mau	4kmt
Эта работа	1.32	2.03	**1.93**	1.52	1.36
RosettaAntibody	**0.55**	**1.89**	2.27	**0.67**	**1.13**

Модель\структура	4m6m	4kuz	4kq3	4m6o	4ma3
Эта работа	**1.57**	**1.61**	1.24	**1.82**	2.9
RosettaAntibody	2.58	2.40	**1.15**	3.9	2.44

Эта работа, среднее: 1.73 Å RMSD
RosettaAntibody, среднее: 1.90 Å RMSD
Датасет AMA-II

Результаты

- Выбраны и адаптированны векторные представления белковых последовательностей
- Предложен способ оптимизации отклонения координат последовательности с сохранением физического смысла
- На основе экспериментов с несколькими архитектурами глубоких нейронных сетей выбрана наиболее подходящая
- Модель на основе CharLSTM позволяет достичь схожих результатов с классическими алгоритмами на датасете AMA-II