Оценка качества прогнозирования структуры белка с использованием графовых нейронных сетей.*

Северилов П.А.[1]
severilov.pa@phystech.edu
[1]Московский физико-технический институт (МФТИ)

Оценка качества предсказания белковой структуры является важной и пока открытой проблемой в структурной биоинформатике (биологии). В работе проводится анализ графовых нейронных сетей в комбинации со сверточными применительно к данной задаче.

Ключевые слова: *белок, графы, графовые нейронные сети, GCN.*

1 Введение

Понимание белковых структур и выполняемых задач помогают контроллировать биологические процессы. Белки спонтанным образом принимают форму в различных средах [?] — форма диктует функционал. Но из имеющихся последовательностей аминокислот в белке трудно определить, в какую форму произойдет сворачивание. Идентификация структуры занимает большое количество времени и ресурсов, к тому же, не всегда возможна.

Вычислительные методы, которые решают задачу предсказания структуры в основном состоят из двух этапов[?]: генерация конформаций белка из их аминокислотных последовательностей и оценивание качества предсказания. В данной работе рассматривается только вторая задача. Данная проблема является крайне важной[?]. Каждые два года проводятся соревнования Critical Assessment of protein Structure Prediction (CASP) по решению этой задачи.

До недавнего времени лучшими методами предсказания стурктуры считались[?...?] объединение подходов, основанных на функциях, предназначенных для узкого класса белков. Методы глубинного обучения превзошли [4] эти результаты.

Основные результаты в этой области полагаются на сверточные нейронные сети (CNN) [3]. Т.к. имеющиеся данные представляют собой трехмерные координаты атомов, то предлагается использовать графовые архитектуры нейронных сетей в комбинации с уже имеющимися архитектурами.

2 Связанные работы

To be done
One of the interesting links

3 Постановка задачи

3.1 Задача регрессии

Пусть $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ — заданная выборка, где $\mathbf{X} \in \mathbb{R}^{m \times n \times 3}$ — тензор объект-признак, объекты $\mathbf{x}_i \in \mathbb{R}^{1 \times n \times 3}, i = \overline{1, m}$ — это молекулы, каждая из которых описана множеством 3-мерных координат всех ее атомов, а $\mathbf{y} = [y_1, \ldots, y_m]^\mathsf{T} \in \mathbb{R}^{m \times 1}$ — оценка близости предсказанной и реальной структуры белка. Оценка близости может быть измерена различными метриками: CAD-score [1], LDDT, GDT. В данной работе выбран CAD-score.

*Научный руководитель: В.В. Стрижов

Рассмотривается множество параметрических моделей $\mathfrak{F}$, взятых из класса графовых сверточных нейронных сетей: $\mathfrak{F} = \{\mathbf{f}_k \colon (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \mid k \in \mathfrak{K}\}$, где $\mathbf{w} \in \mathbb{W}$ – параметры модели, а $\hat{\mathbf{y}} \in \mathbb{R}^{m \times 1}$ - вектор оценок предсказаний (CAD-scores).

Рассматривается задача регрессии для предсказания численного значения CAD-score y_i белка на основе его смоделированной пространственной структуры $\mathbf{x}_i$.

Параметры модели $\mathbf{w} \in \mathbb{W}$ подбираются в соответствии с минимизацией функции ошибки на обучении. Определим функцию ошибки $\mathfrak{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = (\hat{\mathbf{y}} - \mathbf{y})^2$, где $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ – CAD-score, предсказанный моделью $\mathbf{f}$, $\mathbf{y}$ – данный в выборке CAD-score.

3.2 CAD score

Обозначим через G множество всех пар элементов последовательности аминокислот (остатков) (i, j), имеющих ненулевую площадь контакта $T_{(i,j)}$ в реальной структуре. Затем для каждой пары остатков $(i, j) \in G$ вычисляется площадь контакта $M_{(i,j)}$ смоделированной структуры. Для каждой пары остатков $(i, j) \in G$ определяется разность площадей контакта как абсолютная разница площадей контакта между остатками i и j в реальной T и смоделированной структуре M:

$$\mathrm{CAD}_{(i,j)} = \left| T_{(i,j)} - M_{(i,j)} \right|$$

Для вычислительной стабильности берется ограниченный CAD: $\mathrm{CAD}^{\mathrm{bounded}}_{(i,j)} = \min\left(\mathrm{CAD}_{(i,j)}, T_{(i,j)}\right)$. Таким образом: CAD-score для всей структуры определяется как

$$\mathrm{CAD}_{\mathrm{score}} = 1 - \frac{\sum_{(i,j)\in G} \mathrm{CAD}^{\mathrm{bounded}}_{(i,j)}}{\sum_{(i,j)\in G} T_{(i,j)}}$$

4 Теоретическая часть

4.1 Представление белков в виде графов

Элементы аминокислотной последовательности рассматриваются как отдельные узлы, чьи связи (ребра) описывают пространственные отношения между ними.

В общем случае граф $\mathbf{G}$ определяется набором $(\mathbf{V}, \mathbf{A})$, где $\mathbf{V} \in \mathbb{R}^{n \times c}$ определяет вершины или узлы графа, n – число узлов и c – число признаков в каждом узле. Матрица смежности $\mathbf{A} \in \mathbb{R}^{n \times n}$ определяет соединения между n узлами (ребра), где $\mathbf{A}_{ij}$ – сила связи между узлами i и j. Используя это определение графа, белковые структуры можно определить как графы, признаки элементов аминокислотной последовательности которых закодированы в элементах $\mathbf{V}$ узлов, а пространственная близость между элементами закодирована в матрице смежности $\mathbf{A}$.

4.2 Слой свертки графа

Дан граф $\mathbf{A}$ и матрица с информацией об узлах $\mathbf{X} \in \mathbb{R}^{n \times c}$. Слой свертки графа представлен в следующей форме:

$$\mathbf{Z} = f\left(\tilde{\mathbf{D}}^{-1}\mathbf{A}\mathbf{X}\mathbf{W}\right),$$

где $\mathbf{A}$ – матрица смежности графа с добавлением петель, $\tilde{\mathbf{D}}$ это его диагональная матрица степеней вершин, где $\tilde{\mathbf{D}}_{ii} = \sum_j \mathbf{A}_{ij}, \mathbf{W} \in \mathbb{R}^{c \times c'}$ – матрица параметров свертки обучаемого графа, f – нелинейная функция активации, а $\mathbf{Z} \in \mathbb{R}^{n \times c'}$ – выходная матрица.

5 Вычислительный эксперимент

5.1 Данные

Берутся с соревнований CASP. Для реальной структуры белка берется еще смоделированная структура. Для них вычисляется CAD-score. Модель на тесте предсказывает CAD-score для смоделированной структуры, не имея возможности напрямую вычислить CAD-score по реальной структуре.

Пример анализа одного из белков:

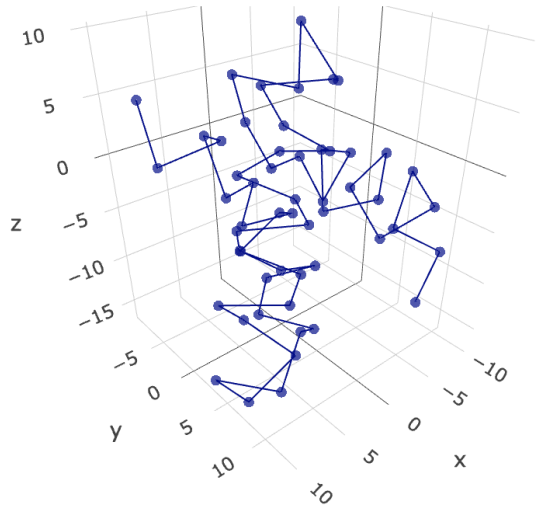

Рис. 1 3D структура белка

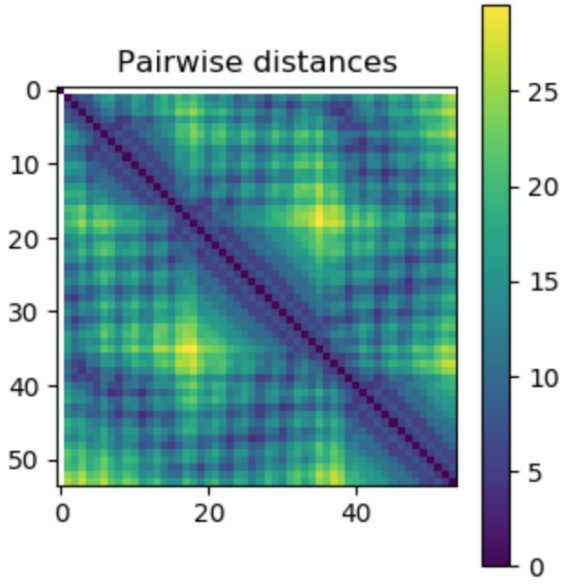

Рис. 2 Попарные расстояния между элементами белка

5.2 Архитектуры сетей

1. Deep Graph Convolutional Neural Network (DGCNN) [6]
2.
3.

6 Результаты

Литература

[1] Kliment Olechnovic, Eleonora Kulberkytė, and eslovas Venclovas. Cad-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins*, 81 1:149–62, 2013.

[2] Angelo Oliveira and Renato José Sassi. Behavioral malware detection using deep graph convolutional neural networks, Nov 2019.

[3] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, 02 2019.

[4] J.Kirkpatrick L.Sifre T.F.G.Green C.Qin A.Zidek A.Nelson A.Bridgland H.Penedones S.Petersen K.Simonyan S.Crossan D.T.Jones D.Silver K.Kavukcuoglu D.Hassabis A.W.Senior R.Evans, J.Jumper. De novo structure prediction with deep-learning based scoring. *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4*, Dec 2018.

[5] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596, 2019.

[6] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. 2018.

[7] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *CoRR*, abs/1812.08434, 2018.