

Deep Learning for reliable detection of tandem repeats in 3D protein structures.*

Веселова Е.Р.¹

veselova.er@phystech.edu

¹Московский физико-технический институт (МФТИ)

В работе рассматривается улучшение алгоритма выделения осей симметрии и повторяющихся элементов в трёхмерной структуре белка. Общая задача анализа трёхмерных белковых структур с высокой эффективностью решается с помощью свёрточных нейросетей (CNN — convolutional neural network). Относительно трансляции данных CNN обладают свойством устойчивости, или *эквивариантности* — коммутативности преобразования исходных данных и свёртки, однако для вращений данное свойство не сохраняется. Обеспечить эквивариантность всей нейросети предлагается заменой классических свёрточных фильтров на линейную комбинацию трёхмерных сферических гармоник, эквивариантных относительно вращений. Решение данной задачи позволит увеличить точность и скорость обработки белков.

Ключевые слова: *CNN, свёрточные нейросети, сферические гармоники, структурные повторения, оси симметрии, 3D объект.*

1 Введение

Машинное обучение широко применяется в задачах современных естественных наук, в частности в задачах структурной биологии. Большая часть получаемых на практике белков обладает повторяющимися элементами структуры или симметрией, которые влияют на функции белков и позволяют исследовать их эволюцию. Нахождение симметрий и повторов является важной задачей, во многих формулировках уже решённой. Решённой с помощью классических методов машинного обучения в 2006 году [4]. При наличии точечного представления плотности 3D объекта задача определения всех типов симметрий была решена аналитически. Поэтому особый интерес представляет применение методов глубокого обучения и свёрточных нейросетей (CNN — convolutional neural network), которые позволяют получать на нижних уровнях сети легко интерпретируемые характерные черты изучаемых объектов.

Предсказание трёхмерной структуры белка по его аминокислотному составу [1] и детектирование структурных повторений и внутренних симметрий с высокой точностью решается свёрточными нейросетями [6], однако существующие свёрточные нейросети не имеют возможности одинаково качественно обрабатывать входные данные при любых поворотах и сдвигах. Основная цель работы состоит в адаптации построенных нейросетей для выявления повторов и симметрий к различным преобразованиям входных данных.

Если трансляция входного объекта при обработке свёрточной нейросетью даёт пропорционально транслированную карту характеристик [3], то для вращений входного объекта подобное свойство не реализуется. Искомое свойство переноса преобразования входных данных на выходные называется эквивариантностью. Кроме того, сохраняющаяся во всех слоях нейросети эквивариантность позволяет отслеживать свойства исследуемых структур уже на нижних уровнях нейросети. 2D CNN, относительно которой данные были бы

*Задачу поставил: Grudinin S. Консультант: Pages G.

эквивариантны, была реализована заменой стандартных свёрточных фильтров на комплексные *круговые гармоники* (circular harmonics), обеспечивающие вращательную эквивариантность без необходимости использовать сильную аугментацию данных [8]. Трёхмерной интерпретацией данного подхода являются сферические гармоники. Первоначально идея сферических гармоник была развита в моделировании для эффективного представления и определения степени схожести поверхностей трёхмерных объектов [2, 5]. Далее идея была применена к анализу трёхмерных признаков карт с помощью CNN. При замене стандартных трёхмерных свёрточных фильтров на линейную комбинацию аналитически определённого вращательного базиса из *сферических гармоник* CNN становится эквивариантна относительно любого преобразования из группы симметрий $SE(3)$ [7].

Любое движение $g \in SE(3)$ представимо как комбинация вращения $r \in SO(3)$ и трансляции $t \in \mathbb{R}^3$. При рассмотрении одного уровня свёрточной нейросети с K трёхмерных признаков карт, соответствие между входом и выходом слоя может быть записано как $f : \mathbb{R}^3 \rightarrow \mathbb{R}^K$. Оператор трансляции выходного векторного поля легко описывается как $t : (x - t) \mapsto x$. Вращение описывается более сложным образом, так как при повороте всей каждый вектор меняет свою позицию и поворачивается с помощью матрицы $\rho(r)$. Поэтому оператор вращения $\pi(r)$ определяется как $[\pi(r)f](x) := \rho(r)f(r^{-1}x)$, где $r^{-1}x$ описывает перемещение векторов на новые позиции. Таким образом, $g = tr$ представимо как $[\pi(tr)f](x) := \rho(r)f(r^{-1}(x - t))$. Обобщая полученные выкладки на CNN, выражение свёрточного фильтра между n и $n + 1$ слоями нейросети через базис в пространстве эквивариантных преобразований между пространствами признаков \mathcal{F}_n и \mathcal{F}_{n+1} позволяет гарантировать, что любое преобразование входа слоя будет давать такое же преобразование выхода, т.е. обеспечивать эквивариантность. Кроме того, в силу одинакового изменения всех трёх RGB матриц изображения при рассматриваемых преобразованиях полученное представление может быть перенесено и на цветные изображения.

Именно поэтому в качестве решения поставленной задачи в статье предложена эффективная имплементация сферических гармоник в существующую CNN модель выделения тандемных повторов и симметрий в белках для получения идентичных результатов при любых вращениях исходных карт атомных плотностей белковых 3D моделей [6]. В качестве входных данных выступает синтетический датасет, полученный «симметризацией» белковых структур датасета Top8000*, состоящий из карт плотностей размеров $24 \times 24 \times 24$.

2 Постановка задачи

Данными в задаче являются $K = 11$ карт атомных плотностей элементов белка. Каждая карта $\mathbf{x}_i \in \mathbb{R}^{24 \times 24 \times 24}$ обучающей выборки \mathbf{X} получена искусственной симметризацией карт датасета Top8000 с порядком циклической симметрии от 1 до $N_{order} \in [10, 20]$. Ответом на элементе выборки $\mathbf{f}(x_i) = \mathbf{f}_i \in \mathbb{R}^{N_{order}+6}$ является композиция двух векторов $\mathbf{y}_i \in \mathbb{R}^{N_{order}}$ и $\mathbf{z}_i \in \mathbb{R}^6$, где \mathbf{y}_i определяет вероятность каждого порядка симметрии, \mathbf{z}_i определяет положение оси симметрии. Размерность 6 для задания оси симметрии выбрана неслучайно: модель переводит ось в трёхмерном пространстве в её представление в шестимерном пространстве посредством *отображения Веронезе* $V(x, y, z) = (x^2, y^2, z^2, \sqrt{2}yz, \sqrt{2}zx, \sqrt{2}xy)$.

Исходная модель $\mathbf{f}(\mathbf{X})$ — свёрточная нейросеть, состоящая из двух свёрточных слоёв, имеющих по 4 фильтра каждый, и трёх полносвязных слоёв, входом для которых является конкатенация столбцов выхода второго свёрточного слоя. Зададим $\mathbf{W} \in \mathbb{R}^{n \times 4 \times 2}$ — матрицу

*<http://kinemage.biochem.duke.edu/databases/top8000.php>

коэффициентов разложения каждого из 4 фильтров двух свёрточных слоёв нейросети на n базисных сферических гармоник. Требуется адаптировать модель $\mathbf{f}(\mathbf{X}, \mathbf{W})$ так, что при воздействии на исходные данные любым оператором $\pi(tr)$ трансляции t и вращения r реализуется свойство эквивариантности, т.е.

$$\mathbf{f}(\pi(tr)\mathbf{X}, \mathbf{W}) = [\pi(tr)\mathbf{f}](\mathbf{X}, \mathbf{W}).$$

Функция потерь при определении порядка симметрии объекта \mathbf{x}_i при известном истинном порядке k_t с помощью вероятности каждого порядка $P(k) = \frac{\exp(p_k)}{\sum_{j=1}^{N_{order}} \exp(p_j)}$ для вектора $\mathbf{y}_i = [p_1, \dots, p_{N_{order}}]^T$ может быть записана как

$$L_c(\mathbf{W}) = -\log(P(k_t)) = \log\left(\sum_{j=1}^{N_{order}} \exp(p_j)\right) - p_{k_t}.$$

Функция потерь при определении оси симметрии объекта \mathbf{x}_i по известным координатам одной из точек (x_t, y_t, z_t) истинной оси симметрии для вектора \mathbf{z}_i может быть записана как

$$L_a(\mathbf{W}) = \|\mathbf{z}_i - V(x_t, y_t, z_t)\|_2.$$

Получаем решаемую в процессе обучения задачу оптимизации

$$\mathbf{W}^* = \arg \min_{\mathbf{W}} (L_c(\mathbf{W}) + L_a(\mathbf{W}))$$

3 Базовый эксперимент

С целью проверки работоспособности предложенной идеи был проведён базовый эксперимент, в ходе которого была реализована структура базовой нейросети из работы [6] с применением свёрточных слоёв из модуля `se3cnn` [7]. Базовая архитектура была переработана с учётом требований модуля и необходимого порядка сферических гармоник. Согласно эмпирическим данным, достаточно второго–третьего порядков сферических для появления эквивариантности в структуре, поэтому свёрточный фильтр в конволюционном слое представляет из себя комбинацию трёх наборов сферических функций порядков 0, 1 и 2 с двумя наборами коэффициентов разложения для каждого из наборов. Таким образом, выход конволюционного слоя имеет $2 * ((2 * 0 + 1) + (2 * 1 + 1) + (2 * 2 + 1)) = 18$ карт. Синтетические данные для обучения тестовой сети представляют из себя 400 симметричных с порядком 3 относительно любой из главных диагоналей матриц. Выходом для тестовой сети является заданная тремя координатами ось симметрии.

| Layer | Type | Input dimensions | Output dimensions | Parameters |
|-------|-----------------|------------------------------------|------------------------------------|------------------|
| 1 | SE3Convolution | $24 \times 24 \times 24 \times 1$ | $21 \times 21 \times 21 \times 18$ | Size 4, stride 1 |
| 2 | Average pooling | $21 \times 21 \times 21 \times 18$ | $8 \times 8 \times 8 \times 7$ | Size 2, stride 2 |
| 5 | Reshape | $8 \times 8 \times 8 \times 7$ | 1512 | |
| 6 | Linear | 1512 | 800 | |
| 7 | LeakyReLU | 800 | 800 | |
| 8 | Linear | 800 | 3 | |

Литература

- [1] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *ArXiv e-prints*, 2018.

-
- [2] Michael Kazhdan and Thomas A. Funkhouser. Harmonic 3d shape matching. In *SIGGRAPH Abstracts and Applications*, page 191. ACM, 2002.
 - [3] Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. *International Journal of Computer Vision*, 2018.
 - [4] Niloy J. Mitra, Leonidas Guibas, and Mark Pauly. Partial and Approximate Symmetry Detection for 3D Geometry. *ACMTG: ACM Transactions on Graphics*, 25, 2006.
 - [5] Mohamed-Hamed Mousa, Raphaëlle Chaine, Samir Akkouché, and Eric Galin. Toward an efficient triangle-based spherical harmonics representation of 3D objects. *Computer Aided Geometric Design*, 25(8):561–575, 2008.
 - [6] Guillaume Pagès and Sergei Grudinin. DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. working paper or preprint, 2018.
 - [7] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *CoRR*, abs/1807.02547, 2018.
 - [8] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. In *CVPR*, pages 7168–7177. IEEE Computer Society, 2017.