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

Веселова Е.Р.[1]
veselova.er@phystech.edu
[1]Московский физико-технический институт (МФТИ)

В работе рассматривается задача регрессионного выделения осей симметрии трёхмерных объектов и классификационного выделения порядков найденных осей. Обе задачи решаются с помощью применения свёрточных нейросетей к синтетическому датасету, полученному размножением 3D моделей белковых структур. Относительно трансляции данных свёрточные нейросети обладают свойством устойчивости, что не выполняется для вращений. Предлагается применение трёхмерных сферических гармоник вместо классических свёрточных фильтров в CNN. Решение данной задачи позволит увеличить точность обработки и автоматического выделения свойств белков.

Ключевые слова: *CNN, сферические гармоники, ось 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 повтояющимися элементами структуры или симметрией, которые влияют на функции белков и позволяют исследовать их эволюцию. Нахождение симметрий и повторов является важной задачей, решённой с помощью классических методов машинного обучения в 2006 году [3], 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 тандемных повторов и внутренних симметрий с высокой точностью решается свёрточными нейросетями [5], 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 [2], 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 эквивариантны, была реализована заменой стандартных свёрточных фильтров на комплексные круговые гармоники (circular harmonics), обеспечивающие вращательную эквивариантность без необходимости использовать сильную аугментацию данных [7]. Трёхмерной интерпретацией данного подхода являются сферические гармоники. Первоначально идея сферических гармоник была развита в моделированиидля более эффективного рендеринга поверхностей трёхмерных объектов [4]. Далее идея была применена к анализу

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

трёхмерных признаковых 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 $SE(3)$ [6].

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 признаковых карт, соответствие между входом и выходом слоя может быть записано как $f : \mathbb{R}^3 \to \mathbb{R}^K$. Оператор трансляции выходного векторного поля легко описывается как $t : x - t \rightarrowtail x$. 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 $[\pi(tr)f](x) := \rho(r)f(r^{-1}(x-t))$. 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 [5]. В качестве входных данных выступает синтетический датасет, полученный «симметризацией» белковых структур датасета Top8000*, состоящий из карт плотностей размеров $24 \times 24 \times 24$.

2 Вывод

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] Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. *International Journal of Computer Vision*, 2018.

[3] Niloy J. Mitra, Leonidas Guibas, and Mark Pauly. Partial and Approximate Symmetry Detection for 3D Geometry. *ACMTG: ACM Transactions on Graphics*, 25, 2006.

[4] Mohamed-Hamed Mousa, Raphaëlle Chaine, Samir Akkouche, and Eric Galin. Toward an efficient triangle-based spherical harmonics representation of 3d objects. *Computer Aided Geometric Design*, 25(8):561–575, 2008.

[5] 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.

[6] 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.

[7] 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.

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