PhD subject

DeepFOLD: Deep Learning for Protein Fold Recognition

Hosting institute

— ICube laboratory: Le laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (The Engineering science, computer science and imaging laboratory), http://icube.unistra.fr/

— University of Strasrboug, France, http://www.unistra.fr/

Work place

Place de l'hôpital, Strasbourg (67), France.

Supervisors

— Claudine Mayer (ICube, Univ. Strasbourg et Univ. Paris Diderot)

— Hyewon Seo (ICube, Univ. Strasbourg)

Staring date

September/October 2020.

Subject description

In the last decade, research in computer vision and artificial intelligence has achieved disruptive results in the recognition and synthesis of objects in images by means of deep learning (DL) algorithms, large annotated datasets, and adequate GPU resources. However, most existing architectures and algorithms having been developed for 2D images, recent successful learning techniques, e.g. convolutional deep networks (CDN), do not easily generalize to the most common form of 3D data (point clouds or polygonal meshes).

While CDNs have been used in some 3D data, e.g. face [TZK+17, TL18] and body modeling [TBC17, TYF17, KBJM18], 3D object [WXX+16, MS15, WSK+15] or scene [LYB19] recognition and classification, their interest in modeling of complex 3D shapes has not yet been fully explored, and extending the learning ability to 3D data remains largely an unexplored area. The majority of DL techniques designed for 3D data uses volume as input [WXX+16, MS15, WSK+15], a straightforward 3D equivalent of the regular grid structure of 2D images. Such volume representation drastically limits the dimension of data that can be learned, and more importantly, does not describe well the boundary shape as found in most surface data, i.e. triangle mesh. Alternatively, deep neural networks (DNN) developed for estimating 3D shapes from 2D image input adopt limited 3D shape representation such as depth map [SSN09], UV map [GNK18], or parametric models of known class of objects such as face [TZK+17, TL18] or body [TBC17, TYF17, KBJM18].

The aim of this thesis study is to represent and learn 3D structure of a specific class of biological objects, e.g. proteins. Determining and studying 3D structure of proteins is a key question in structural

biology, because structure is related to biological function. Each 3D structure can be described by a defined topology, e.g fold. Interestingly, the total number of folds in nature is limited to nearly 1500, leading to the relevant question of how protein fold and biological function are related. **More specifically, we will design and develop devoted descriptors as well as DL architecture that can identify and characterize the fold based on the tertiary (3-dimensional) structure of a protein chain, which is composed of a defined sequence of amino acids.**

Previous attempts have used sequence-conservation or frequency-based descriptors of primary [DD01] and secondary [WLGZ15, ZDW11] structures, neglecting 3-dimensional structural features. Recent DL techniques developed in this context tend to focus on fold recognition by using template protein whose 3D structure is known [JHEC15], or sequence-to-fold prediction [HAC18]. However, most of them deploy variants of existing descriptors, and fewer exist that can consume directly 3D structure as input.

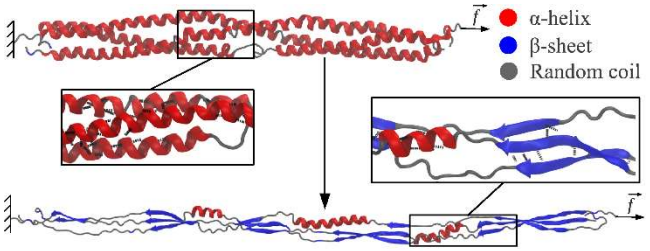

Figure 1: Secondary structure of proteins: Helix, strand, and coil [ZKL+12].

Arguably, it is of great interest to be able to analyze protein folds by directly working on the 3D shape. By fully exploiting the 3D data (rather than only 1D and/or 2D as with conventional sequence-based or secondary structure assignment descriptors), we can expect to profoundly improve the performance of DL techniques in many relevant tasks, such as recognition, classification, even synthesis of targeted proteins.

We are motivated by the observation that the 3D protein structure exhibits a unique characteristic: a few known types of wire shapes (defined as secondary structures called alpha-helix, beta-sheet, or coil: see *Figure 1*) "folded" in a relatively compact 3D space. This suggests that efficient, devoted shape descriptors as well as DL architecture can be developed, unlocking the limited use of recent learning techniques in studying proteins, fundamental structural and functional elements within every living cell.

As a first step, we will investigate on new, specific descriptors efficiently encoding the tertiary structure of proteins or protein domains, which is composed of a single polypeptide chain "backbone" with one or more protein secondary structures. To describe the global 3D structure, we propose to adopt graph, which can be consumed by very recent DNNs such as GCN (Graph convolutional network) [VBV18]. The sequential information as well as the spatial arrangement can be coded by using a series of 3-dimensional vectors in the form of node/edge attributes. Next, we will develop an optimal DL architecture for the new descriptors. The possibility of combining the GCN architecture with an RNN (Recurrent neural network) [Grav11] will be studied, the latter being designed to learn sequential data. Finally, we will test and tune hyper parameters of the network, and compare its performance with existing state-of-the-art methods. We note that there exist several public datasets for the training/test for the network: PDB (Protein Data Bank), SCOPe (Structural Classification of Proteins — extended), and CATH.

Requirements

— Master degree in Computer Science

— Working knowledge of programming in Python/Matlab

— Experience in machine learning

— Knowledge in structural biology is a plus

Bibliography

[Chou05] K.-C. Chou, « Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes », Bioinformatics, vol. 21, no 1, p. 10-19, janv. 2005.

[DD01] C. H. Ding et I. Dubchak: Multi-class protein fold recognition using support vector machines and neural networks. Bioinforma. Oxf. Engl., vol. 17, no 4, p. 349-358, avr. 2001.

[GNK18] Riza Alp Güler, Natalia Neverova, Iasonas Kokkinos: DensePose: Dense Human Pose Estimation in the Wild. CVPR 2018: 7297-7306.

[Grav11] Supervised Sequence Labelling with Recurrent Neural Networks, Alex Graves, Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

[HAC18] Hou, Jie & Adhikari, Badri & Cheng, Jianlin. (2018). DeepSF: Deep Convolutional Neural Network for Mapping Protein Sequences to Folds. 565-565. 10.1145/3233547.3233716.

[ICNK17] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Deep learning advances in computer vision with 3d data: A survey. CSUR, 2017.

[JHEC15] Jo, Taeho & Hou, Jie & Eickholt, Jesse & Cheng, Jianlin. (2015). Improving Protein Fold Recognition by Deep Learning Networks. Scientific Reports. 5. 17573. 10.1038/srep17573.

[KBJM18] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR. (2018)

[LYB19] Liu, Xingyu & Yan, Mengyuan & Bohg, Jeannette. (2019). MeteorNet: Deep Learning on Dynamic 3D Point Cloud Sequences.

[MS15] D. Maturana and S. Scherer. 2015. VoxNet: A 3D convolutional neural network for real-time object recognition. In IEEE/RSJ International Conference on Intelligent Robots and Systems. 922–928.

[SSN09] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. 2009. Make3D: Learning 3D Scene Structure from a Single Still Image. IEEE Trans. Pattern Anal. Mach. Intell. 31, 5 (May 2009), 824-840.

[TBC17] Tan, V., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3D human body shape and pose prediction. In: BMVC. (2017)

[THMM17] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3d morphable models with a very deep neural network,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[TL18] L. Tran, X. Liu: Nonlinear 3D Face Morphable Model. IEEE Conf. IEEE Conf. Computer Vision and Pattern Recognition, 2018

[TYF17] Tung, H., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NIPS. (2017)

[TZK+17] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, C. Theobalt: MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. Int’l Conf. Computer Vision 2017.

[WXX+16] Y.Wang, Z. Xie, K. Xu, Y. Dou, and Y. Lei. 2016. An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning. Neurocomputing 174 (2016), 988–998.

[WSK+15] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 2015. 3D shapenets: A deep representation for volumetric shapes. In IEEE Conference on Computer Vision and Pattern Recognition. 1912–1920.

[WLGZ15] L. Wei, M. Liao, X. Gao, et Q. Zou, « Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique », IEEE Trans. Nanobioscience, vol. 14, no 6, p. 649-659, sept. 2015.

[ZDW11] S. Zhang, S. Ding, et T. Wang, « High-accuracy prediction of protein structural class for low similarity sequences based on predicted secondary structure », Biochimie, vol. 93, no 4, p. 710-714, avr. 2011.

[ZKL+12] Artem Zhmurov, Olga Kononova, Rustem I. Litvinov, Ruxandra I. Dima, Valeri Barsegov, John W. Weisel. Mechanical Transition from α-Helical Coiled Coils to β-Sheets in Fibrin(ogen), J. Am. Chem. Soc. 2012, 134, 50, 20396-20402