SORBONNE
UNIVERSITÉ

PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

Intitulé du projet de recherche doctoral (PRD):
Geometric deep manifold learning combined with NLP for protein movies

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : LAINE Prénom : Elodie

Titre : Maître de Conférences (HDR)

e-mail : elodie.laine@sorbonne-universite.fr

Adresse professionnelle : 4, place Jussieu, 75005 PARIS
(site, adresse, bât., bureau)

Unité de Recherche :

Intitulé : Laboratoire de Biologie Computationnelle et Quantitative (LCQB)

Code *(ex. UMR xxxx)* : UMR 7238, CNRS-Sorbonne Université

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **EDITE**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 1, 2019, 50%

--

Co-encadrant.e :

NOM : GRUDININ Prénom : Sergei

Titre : CRCN CNRS HDR dispense de 3 ans

e-mail : sergei.grudinin@inria.fr

Unité de Recherche :

Intitulé : Laboratoire Jean Kuntzmann (LJK)

Code *(ex. UMR xxxx)* : UMR 5224, Inria Grenoble Rhône-Alpes - CNRS

École Doctorale de rattachement :

Ou si ED non Alliance SU : MSTII Grenoble

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 3, 2017-80% (défense prévue en mars 2021), 2020-100%, 2019-20%

Cotutelle internationale : Non Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
Oui, Les aspects de modélisation mathématique, de développement d'algorithmes efficaces, de traitement de données à large échelle, et d'application à la biologie sont susceptibles d'intéresser l'ISCD.

Description du projet de recherche doctoral *(en français ou en anglais) :*

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Merci d'enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l'initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.

Geometric deep manifold learning combined with NLP for protein movies

Background

Artificial intelligence, and more specifically deep learning, has recently emerged as a powerful approach to exploit the massive amounts of protein sequence and structure data available nowadays toward guiding biological intervention to improve human health. A couple of months ago, the alphaFold2 architecture from DeepMind revolutionised the field of protein structure prediction by reaching unprecedented levels of near-experimental accuracy [1]. This achievement has been made possible mostly thanks to the latest improvements in geometric learning and natural language processing (NLP) techniques. In parallel, several groups have shown that knowledge can be effectively transferred from semi-supervised learning of huge amounts of (meta-)genomics data to a wide range of supervised downstream tasks, including protein classification, functional annotation, mutational outcome prediction and contact inference [2-3].

While the problem of determining how a protein folds in three dimensions (3D) is essentially solved, accessing protein motions is becoming more central than ever before. At the European level, the ELIXIR community is investing efforts right now to create a comprehensive resource for structural diversity and flexibility in the Protein Data Bank (PDB) [8], which contains all experimentally-determined protein 3D structures. Indeed, proteins are flexible biological objects, constantly moving and changing their shape to interact with their environment and cellular partners. This inherent flexibility is highly relevant for protein functioning [4]. For instance, single mutations or small sequence variations may have dramatic consequences on the ability of proteins to adapt to their partners, without any visible effect on their static 3D structure [5-6]. Moreover, targeting protein motions is a powerful strategy to control protein function in physio-pathological contexts and develop drugs with less side effects [7]. Experimentally, it is very difficult to observe proteins directly in action, and we have mostly access to isolated clusters of "snapshots" (conformations) representative of a few functional states. Biomolecular simulations can be used to generate conformational ensembles, but they remain computationally costly. Alternatively, relatively simple physics-based models where the protein is represented by an elastic network have proven very useful to extrapolate functional motions, starting from a single structure [8-9]. Nevertheless, these models are unable to capture changes involving substantial rearrangements in the topology of the starting network.

Objectives

The goal of this project is to explore the contribution of recent methods of statistical learning and deep neural networks to predict motions and conformational states relevant to protein functioning. In other words, **we aim at learning low-dimensional motion manifolds using sparse high-dimensional observations (3D structures)**. Our specific objectives will be to:

1. Develop algorithms capable of operating on compact representations of geometric data structures, taking into account the specific (physico-chemical) constraints applying on them and the uncertainties in the observations;
2. Develop a generative model able to recover observed states, interpolate between them and extrapolate to previously unseen states. The generation will be conditioned on the protein sequence of amino acids, and on low-resolution experimental data for guiding the extrapolation;
3. Generate new plausible states for a set of proteins with therapeutic interest, that could be targeted by small molecules toward modulating their function.

Data

As input data, we will use the protein 3D structures contained in the PDB [10], and also 3D conformations generated from elastic networks representing these structures [8-9]. We will focus on specific protein families retrieved from the Pfam database [11], *e.g.* K-Ras/H-Ras, TSG/OG, Active/Inactive Kinase, which are therapeutic targets and for which abundant structural information is available. To validate the predictions, we will use experimental structures, cryo-electron-microscopy movies and data generated by biomolecular simulations (MoDEL [12], GPCRMD [13]...). Low-resolution solution experimental data collected at physiological conditions, such as small-angle scattering profiles (SAXS), will be used as an additional validation and as priors for the generation of previously unseen states. We would like to emphasise that the proteins' environments (pH, ligands, partners...) are reflected in the experimental observations, and hence, they will be implicitly modelled.

Methodology and preliminary work

Motion extraction from a set of static molecular structures can be seen as manifold learning [14]. Our working hypothesis will be that 3D molecular shapes, albeit highly complex, lie on learnable low-dimensional manifolds. This is supported by the observations that the number of distinct protein functional states is very limited [15-17], and by our recent findings that protein functional motions can be very efficiently described with only a few nonlinear collective coordinates [7,8]. Specifically, the candidate will build an encoder-decoder architecture that will learn a continuous *K*-dimensional protein motion manifold from a set of static structures defined in a *N*-dimensional space, with *K<<N*. To generate new structures, we will draw samples from the learnt manifold. Since the input 3D data are very sparse, learning the manifold using only geometrical information will be very challenging. To cope with this issue, we will transfer knowledge from publicly accessible transformer models pre-trained on hundreds of millions of sequences observed in nature [2-3]. The attention filters implemented in these models capture long-range dependencies between the protein amino-acid residues, and some of these dependencies are very strong indicators for contacts in 3D [18]. We hypothesise that the information encoded in these filters is also relevant to the way residues move together, and can be exploited to define "basis sets" to reconstruct the motion manifold.

The candidate will investigate different types of representations for the input data, building on and extending recent developments by SG's team to produce locally equivariant representations of protein structures (oriented sets of Gaussian clouds [19], Voronoi 3D tessellations and molecular graphs [20]) and/or rotation-equivariant convolutional operators [21]. We will develop strategies to deal with uncertainties and dependencies in the input data. Specifically, we will explicitly use structural annotations of flexibility and disorder (high temperature factors, missing regions…) as indicators of the "dynamical potential" of the different regions of the input structure. Moreover, we will include some probabilistic account of the dependencies between the experimental observations and the simulated data generated from them. Finally, we will explicitly enforce biological and physical priors in the designed architecture to control its computational and memory footprint and to ease its interpretability.

Expected outcomes

- Novel deep architectures that will combine elements from NLP and geometric learning
- A set of algorithms operating on 3D shapes defined with specific physico-chemical constraints
- A set of family-specific expressive low-dimensional manifolds for protein 3D shapes and motions
- A set of therapeutic targets, in the form of motions and previously uncharacterized conformational states for several disease-related proteins

Suitability for SCAI

The project is highly interdisciplinary, at the interface between computer science, biology, mathematics, and physics. Specifically, it lies at the cross-talk between artificial intelligence, genomics and structural bioinformatics. The project embeds original concepts about the sequence-structure-dynamics-function relationship. It comprises a development part aiming at producing new algorithms for manifold learning combining concepts and techniques from geometric learning and natural language processing, and an applicative part toward the description of protein motions and states that are not accessible experimentally and that could be exploited to develop new drugs.

Role of each supervisor / skills provided

EL and SG will co-supervise the student (50/50). EL will contribute with her expertise in the analysis and manipulation of protein sequences. She has developed methods exploiting the evolutionary relationships between natural sequences for the prediction of protein interactions and mutational outcomes. She has also investigated the impact of single mutations and of alternative-splicing-induced sequence variations on protein structural dynamics. Additionally, she has contributed a proof-of-concept of the targeting of protein motions by drugs.

Previous works in the team of EL directly related to the subject:

- Ait-hamlat A., DJ. Zea, A. Labeeuw, L. Polit, H. Richard and **E. Laine**. (2020) Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the JNK family. *J Mol Biol* 432:2121-2140.
- **Laine E**., Y. Karami and A. Carbone. (2019) GEMME: a simple and fast global epistatic model predicting mutational effects. Mol Biol Evol. 36:2604–2619.
- Karami Y., T. Bitard-Feildel, **E. Laine** and A. Carbone. (2018) "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations, *Scientific Reports*. 8 :16126.
- **Laine E**., C. Goncalves, J. Karst, A. Lesnard, S. Rault, W.-J. Tang, TE. Malliavin, D. Ladant and A. Blondel. (2010) Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor. *PNAS* 107:11277-82.

SG will bring his expertise in machine learning applied to proteins, and more specifically in geometric deep learning. His very unique expertise in the development of deep learning architectures for 3D structures of molecules is recognised both at the national level (invitation to present DL for structural biology at the prospective colloquium « Science des données, IA et biologie », 12/2020) and at the international level (invitation to animate the CASP14 round table on DL, 12/2020, top results in CASP protein structure prediction challenges). He also leads the data-related work-package of the ELIXIR initiative on charting the experimentally sampled conformational diversity of native proteins by exploiting data from the PDB.

Previous works in the team of SG directly related to the subject:

- Igashov, I., Pavlichenko, N., & **Grudinin, S**. (2020). Spherical convolutions on molecular graphs for protein model quality assessment. *arXiv preprint arXiv:2011.07980*.
- Igashov, I., Olechnovic, K., Kadukova, M., Venclovas, C., & **Grudinin, S**. (2021). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. *Bioinformatics*. In press.
- Pagès G, Charmettant B, **Grudinin S**. (2019) Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*. 35:3313-3319.
- Derevyanko G, **Grudinin S**, Bengio Y, Lamoureux G (2018) Deep convolutional networks for quality assessment of protein folds. Bioinformatics. 34:4046-4053.
- Hoffmann, A., & **Grudinin, S**. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *Journal of chemical theory and computation*, *13*(5), 2123-2134.

Both EL and SG have some expertise in the analysis and modelling of protein structures and motions. In the past, they have collaborated to develop approaches combining sequence- and structure-based information to predict protein structures and complexes (joint participation to CASP and CAPRI), and on the prediction/ description protein functional transitions:

- **Grudinin, S**., **Laine, E**., & Hoffmann, A. (2020). Predicting protein functional motions: an old recipe with a new twist. *Biophysical journal*, *118*(10), 2513-2525.
- **Laine, E**., & **Grudinin, S**. (2021). HOPMA: Boosting protein functional dynamics with colored contact maps. *bioRxiv*, 2020-12.

Profile of the desired student:

S/he should have a solid background in computer science or applied mathematics, very good programming skills (C++ and Python) and deep knowledge in linear algebra. S/he should have some knowledge in biology and some familiarity with biological objects such as protein sequences and structures. Teamwork and excellent communication skills are essential for the achievement of the project.

References

1. Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*.
2. Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., ... & Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 622803
3. Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., ... & Rost, B. (2020). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *arXiv preprint arXiv:2007.06225*.
4. Gerstein, M., & Echols, N. (2004). Exploring the range of protein flexibility, from a structural proteomics perspective. *Current opinion in chemical biology*, *8*(1), 14-19.
5. Karami, Y., Bitard-Feildel, T., Laine, E., & Carbone, A. (2018). "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations. *Scientific reports*, *8*(1), 1-18.
6. Ait-Hamlat, A., Zea, D. J., Labeeuw, A., Polit, L., Richard, H., & Laine, E. (2020). Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the JNK family. *Journal of molecular biology*, *432*(7), 2121-2140.
7. Laine, E., Goncalves, C., Karst, J. C., Lesnard, A., Rault, S., Tang, W. J., ... & Blondel, A. (2010). Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor. *Proceedings of the National Academy of Sciences*, *107*(25), 11277-11282.
8. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235-242.
9. Hoffmann, A., & Grudinin, S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *Journal of chemical theory and computation*, *13*(5), 2123-2134.
10. Grudinin, S., Laine, E., & Hoffmann, A. (2020). Predicting protein functional motions: an old recipe with a new twist. *Biophysical journal*, *118*(10), 2513-2525
11. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic acids research*, *32*(suppl_1), D138-D141.
12. Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., ... & Orozco, M. (2010). MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, *18*(11), 1399-1409.
13. Rodríguez-Espigares, I., Torrens-Fontanals, M., Tiemann, J. K., Aranda-García, D., Ramírez-Anguita, J. M., Stepniewski, T. M., ... & Selent, J. (2020). GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nature Methods*, *17*(8), 777-787
14. Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, *1*(2000), 32.
15. Wei, G., Xi, W., Nussinov, R., & Ma, B. (2016). Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chemical reviews*, *116*(11), 6516-6551
16. Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., ... & Fraser, J. S. (2015). Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *Elife*, *4*, e07574.
17. Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., ... & Jackson, C. J. (2016). The role of protein dynamics in the evolution of new enzyme function. *Nature chemical biology*, *12*(11), 944-950.
18. Rao, R., Ovchinnikov, S., Meier, J., Rives, A., & Sercu, T. (2020). Transformer protein language models are unsupervised structure learners. *bioRxiv*.
19. Pagès, G., Charmettant, B., & Grudinin, S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*, *35*(18), 3313-3319.
20. Igashov, I., Olechnovic, K., Kadukova, M., Venclovas, C., & Grudinin, S. (2020). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. *bioRxiv*.
21. Igashov, I., Pavlichenko, N., & Grudinin, S. (2020). Spherical convolutions on molecular graphs for protein model quality assessment. *arXiv preprint arXiv:2011.07980*.