Campagne 2020 Contrats Doctoraux Instituts/Initiatives

Proposition de Projet de Recherche Doctoral (PRD)

Appel à projet ISVI - Intitiative Sces du vivant ses interfaces 2020

Intitulé du Projet de Recherche Doctoral : Data-driven generative modeling of protein sequence landscapes

Directeur de Thèse porteur du projet (titulaire d'une HDR) :

NOM : **WEIGT** Prénom : **Martin**
Titre : Professeur des Universités ou

e-mail : martin.weigt@upmc.fr
Adresse professionnelle : Campus Pierre & Marie Curie, bât C, 4 place Jussieu, 75005
(site, adresse, bât., bureau) Paris, France

Unité de Recherche :
Intitulé : LCQB – Computational and Quantitative Biology
Code *(ex. UMR xxxx)* : UMR 7238

Ecole Doctorale de rattachement de l'équipe & d'inscription du doctorant : **ED130-EDITE**

Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) : 2:

Kai Shimagaki (inscrit à l'EDITE en Oct. 2018 après une année d'Erasmus+ dans l'équipe)

Maureen Muscat (inscrite à l'EDITE en Oct. 2018 après stage M2 dans l'équipe)

Co-encadrant :

NOM : **ZAMPONI** Prénom : **Francesco**
Titre : Directeur de Recherche ou HDR ☒
e-mail : francesco.zamponi@ens.fr

Unité de Recherche :
Intitulé : LPENS – Laboratoire de Physique de l'ENS
Code *(ex. UMR xxxx)* : UMR 8023

Ecole Doctorale de rattachement : **ED564-Physique en IdF** Ou si ED non Alliance SU :

Doctorants actuellement encadrés par le co-directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) : 1:

Dhruv SHARMA - EDPIF - Inscrit en 2017, soutenance prévue en octobre 2020

Cotutelle internationale : ☒ Non ☐ Oui, précisez Pays et Université :

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

3 pages maximum – interligne simple – Ce texte sera diffusé en ligne

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.

Summary (for diffusion):

Proteins belong to the most fascinating complex systems in nature. Playing a crucial role in almost all biological processes, they attract considerable attention at the interfaces of biology, physics, and computer science. Thanks to the sequencing revolution in biology, protein sequence databases have been growing exponentially over the last years. Data-driven modeling approaches, which in recent times increasingly include methods from Artificial Intelligence, are therefore becoming more and more popular in exploring this emergent data richness.

In our doctoral project we suggest to construct highly accurate, generative but interpretable models for protein sequence landscapes by leveraging rapidly expanding sequence databases, inverse statistical physics and deep learning. The landscapes describe the sequence variability in protein families, i.e. ensembles of proteins having common ancestry in evolution, sharing very similar three-dimensional structures and biological functions, but having highly variable amino-acid sequences. To build these models, we will systematically explore generative modeling approaches, ranging from parsimonious but easily interpretable models (e.g. Boltzmann machines, restricted Boltzmann machines) to more powerful, but also less easily interpretable deep generative models (e.g. autoregressive models, variational auto-encoders and generative adversarial networks). We will explore integrative modeling strategies, which combine publicly available sequence data with more quantitative data (deep-mutational scanning) generated by our close collaborator Dr. Olivier Tenaillon (DR INSERM at Bichat).

Uncovering the patterns of natural sequence variability using generative models will allow us to address biological questions of prime importance, including the assessment of mutational effects in proteins (important e.g. in predicting pathological mutations or evolution of drug resistance) and the data-driven design of new protein sequences.

Project description: Please see next two pages for the detailled project description

Merci de nommer votre fichier pdf :
«ACRONYME de l'institut/initiative_2_NOM Porteur Projet_2020 »

à envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 30 mars.

Data driven generative modeling of protein sequence landscapes

A PhD project proposed by **Martin WEIGT** (LCQB SU) and **Francesco ZAMPONI** (LPENS)

Context – Proteins belong to the most fascinating complex systems in nature. They are simultaneously robust and fragile. Being primarily linear heteropolymers, they fold into well defined 3D structures underlying their biological functionality. However, the folds are thermodynamically marginally stable: increasing temperature by a few degrees may denature proteins. Proteins conserve structure and function throughout evolution, in many cases even from bacteria to humans, while substituting 70-80% of their amino acids; however, very few random mutations may destabilize a protein or interrupt its function.

In classical biophysics, proteins are modeled using fine-tuned models of the physical interactions between amino acids [1,2]. Molecular dynamics simulations help to understand folding; Monte-Carlo sampling allows for exploring a protein's 3D conformational space. Deep insight has been gained, but important limitations persist when treating medium-to-large proteins (>100 amino acids) or relevant time scales (>1µs) due to the high computational cost of these simulations.

Biology is currently undergoing a deep transformation to a data-rich science. A fascinating alternative increasingly gains interest in the community at the interface of biology, machine learning, and statistical physics: instead of modeling proteins from first principles and via expensive simulations, **we consider sequence data as a testimony of the unknown rules relating protein sequence to structure, function, and generated by the partially unknown dynamics of protein evolution** [3].

Specific aims – We aim at learning **data-driven generative protein sequence landscapes** – functions relating protein sequence to protein fitness [4] or related phenotypes, cf. the schematic Fig. 1a – directly from rapidly accumulating sequence data [5]. We will interpret the models in terms of structural and functional constraints, and explore them by modeling evolution in landscapes. To this aim, we will combine **state-of-the art methods in artificial intelligence and inverse statistical physics** [6], and integrate public databases with novel data resulting from top-end experimental techniques (gene synthesis, deep-mutational scanning) provided by our experimental collaborator Dr. Olivier Tenaillon (INSERM Paris), which are able to quantitatively characterize thousands of protein variants.

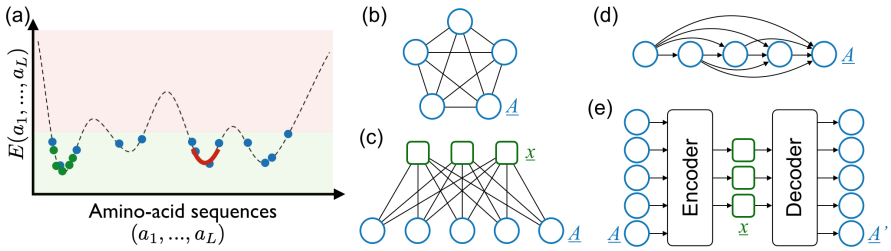

Fig. 1 – (a) Schematic representation of a sequence landscape: Following an analogy to energy landscapes, good (resp. bad) sequences have low (resp. high) values $E(a_1,\ldots,a_L)$. The different data types are shown: blue dots for global landscape sampling by diverged homologs, green dots for local landscape sampling from the intra-species variability, and a red line for a local landscape quantification. Only for the last datatype, $E(a_1,\ldots,a_L)$ is explicitly measured. **(b-e) Architectures of generative models for protein sequences:** The different modeling approaches, to be explored in the proposal, are (b) Boltzmann machines / Potts models, (c) restricted Boltzmann machines / Hopfield-Potts models, (d) autoregressive models and (e) variational auto-encoders. Amino acid variables A are shown as blue circles, latent variables x as green squares. In (b,c) black lines denote pairwise interactions; arrows in (d,e) may represent general dependencies, realized via deep neural architectures.

Data – Our modeling is strongly data-driven, and part of the data will be generated by the project. We will base our work on three data types of different relations to the concept of sequence landscapes, cf. Fig 1a:

A. **Global landscape sampling**: Homologous proteins from different species show, as mentioned before, strong sequence divergence but highly similar collective properties (3D structure and function). Thanks to next-generation sequencing, databases offering this kind of data literally explode. The Pfam protein family database, e.g., lists >17 000 protein families, many with 10^3-10^6 sequences [5].

B. **Local landscape sampling**: Thanks to sequencing thousands of individuals or strains of one species, massive data of almost identical sequences (>90% ID) are available. M. Weigt's team at LCQB has curated a dataset of >60,000 strains of Escherichia coli, and matched with the global Pfam sampling.

C. **Local landscape quantification**: Experiments like deep-mutational scanning [7] allow to quantify the phenotype of thousands of mutants of a protein (e.g. all single-site amino-acid mutations), thus directly measuring the landscape. O. Tenaillon has recognized expertise in generating this type of data [8].

All three data types will be used in the project: many are available in sequence databases, but in particular the more quantitative data related to point C will be generated by our collaborator O.Tenaillon.

Scientific tasks and objectives – Our main objective is to develop data-driven modeling approaches, to test them using accurate quantitative data (published or produced by OT), and to refine them using these data. Particular attention will be laid to the generative character of our landscapes – artificially sampled sequences should be almost indistinguishable from natural ones.

Task 1 – Parsimonious and interpretable models: MW and other groups previously showed that Boltzmann machines [9] (BM, Fig 1b) and Restricted Boltzmann machines [10,11] (RBM, Fig 1c) are interesting candidates for generative models based on inverse statistical physics. However, these models suffer from overfitting due to huge parameter numbers inferred from limited data. We will develop parameter-reduced models, which need less data while remaining generative. RBM contain latent variables allowing for learning dimensionally reduced representations of the sequence data; a shallow architecture makes them more interpretable compared to the deep networks of Task 2.

Task 2 – Deep generative models: Deep learning is at the basis of the recent success of artificial intelligence. While the standard setting is a supervised one, in our context less explored unsupervised approaches are of relevance, such as autoregressive models (ARM, Fig 1d), variational autoencoders (VAE, Fig 1e) or generative adversarial networks (GAN). While first promising results were found for predicting mutational effects [12], their generative capabilities for sequence ensembles remain unexplored. We will analyse the different generative modeling strategies, to establish the currently unknown optimal strategy.

Task 3 – Integrative modeling using all the data: While the first two tasks use only data of type A, we expect further improvements by integrating all data types A, B and C, using in particular the newly generated data by OT, and building upon ideas proposed by MW [13].

All tasks rely on a combination of the experience of MW in modeling biological sequences, and of the expertise of FZ in modeling complex energy landscapes and dynamics [14,15]. MW and FZ have recently started a collaboration at the interface of learning and data-driven biological modeling, co-supervising a postdoc, Anna-Paola Muntoni, now researcher at Politecnico Turin, Italy; 2 papers are in preparation.

References:

[1] Dill, K. A., MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, *338*(6110), 1042-1046

[2] Onuchic, J. N., Wolynes, P. G. (2004). Theory of protein folding. *Current Opinion Struct Biology*, *14*(1), 70-75

[3] Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R., Weigt, M. (2018). Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, *81*(3), 032601

[4] De Visser, J.A.G., Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Rev Gen*, *15*(7), 480

[5] El-Gebali, ... Sonnhammer, E. L. L. (2018). The Pfam protein families database in 2019. *Nucl Acids Res*, *47*(D1), D427-D432

[6] Nguyen, Zecchina, Berg (2017) Inverse statistical problems: from the inverse Ising problem to data science. *Adv Phys 66*, 197

[7] Fowler, D. M., Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, *11*(8), 801

[8] Jacquier, H., ... Tenaillon, O. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. *PNAS*, *110*(32), 13067

[9] Figliuzzi M, P Barrat-Charlaix, and M Weigt. "How pairwise coevolutionary models capture the collective residue variability in proteins?." *Molecular Biology and Evolution* 35.4 (2018): 1018-1027

[10] Tubiana, J., Cocco, S., & Monasson, R. (2019). Learning protein constitutive motifs from sequence data. *eLife*, *8*, e39397

[11] Shimagaki, K., Weigt, M. (2019). Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Physical Review E*, *100*(3), 032128

[12] Riesselman, Adam J., John B. Ingraham, and Debora S. Marks. "Deep generative models of genetic variation capture the effects of mutations." *Nat. Methods* 15 (2018): 816-822

[13] Barrat-Charlaix, P., Figliuzzi, M., & Weigt, M. (2016). Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. *Scientific reports*, *6*, 37812.

[14] Charbonneau, P.,... Zamponi, F. (2014). Fractal free energy landscapes in structural glasses. *Nature Comm*, *5*, 3725.

[15] Franz, S., … Zamponi, F. (2017), Universality of the SAT-UNSAT (jamming) threshold in non-convex continuous constraint satisfaction problems, *SciPost Physics,* 2, 019