Enhanced coevolution signals on protein sequences, with applications to folding and docking

Supervisors : E. Sarti, F. Cazals *
Algorithmes et Biologie Structurale; Inria Sophia Antipolis - Méditerranée †

Context. While determining the amino acid (a.a.) sequence of a protein is experimentally straightforward, determining its 3D structure remains challenging. The in silico determination of structures is a critical research topic, with applications to fundamental biology and medicine.

The sequence of a protein harboring a given function varies slightly across species. The mutations incurred are not random, but instead driven by natural selection: such mutations indeed preserve (and possibly improve) the stability of the protein and the specificity of its interactions. Mutations actually compensate one another, such correlations defining the *coevolution* of amino acids. Conversely, the study of coevolution amidst a set of related protein sequences provides invaluable insights on a.a. accounting for the structure and function of these proteins (Fig. 1).

The coevolution signal can also be used to predict the structure of a protein from its sequence, as evidenced by the ground-braking success of Alphafold by Deepmind [1]. It is also instrumental to predict the interface between two proteins.

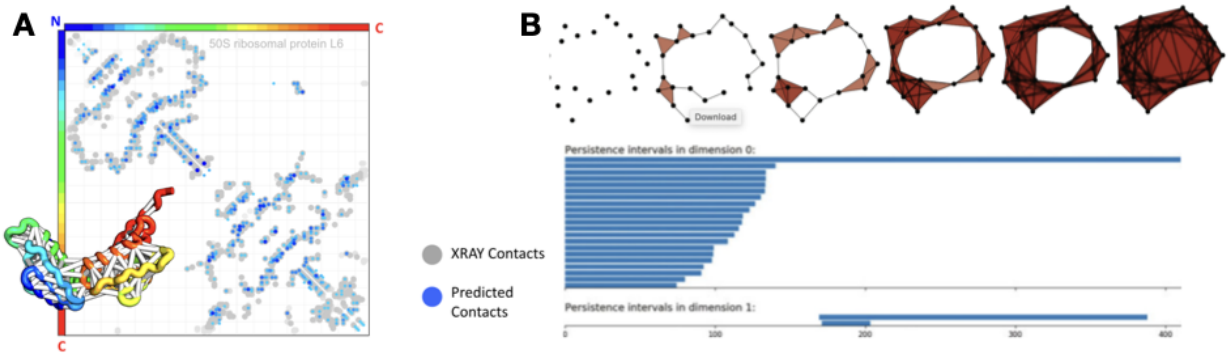

Figure 1: **(A)** Coevolution makes it possible to predict contacts between amino acids which are distant on the sequence [2]. **(B)** Given a sequence of nested *shapes*, persistent homology identifies stable features (connected components, loops, voids) [3].

Objectives. Models derived from coevolution are in general not sparse, with a number of false positive contacts. The goal of this internship will be to derive sparse models derived from coevolution. A key tool to do so will be persistent homology (PH) [3]. In general, PH makes it possible to identify stable structures in an evolving graph / (simplicial) complex (Fig. 1). In our context, it will be used to study the stability of the coevolution signal between a set of amino-acids, with applications both to the prediction of individual structures and of interfaces in protein complexes.

*Emails: edoardo.sarti@inria.fr, frederic.cazals@inria.fr
†http://team.inria.fr/abs

This project will uses algorithms from the Structural Bioinformatics Library (`http://sbl.inria.fr` [4], as well as the direct coupling analysis (DCA) method [5].

Coding skills in python and/or C++ are expected.

Conditions. Internship with gratification. Possibility to follow-up with a PhD thesis.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[2] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030, 2014.

[3] Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule. *Topological Data Analysis*. Springer, 2020.

[4] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[5] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[6] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *arXiv preprint arXiv:2103.03292*, 2021.

[7] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein-protein binding interfaces. *PNAS*, 102(1):57–62, 2005.