Synergizing data-driven models and high-throughput experiments to understand protein function and evolution

PhD and postdoctoral positions

Martin Weigt and Francesco Zamponi

What makes an amino-acid sequence a well-behaved protein? How are protein structure and functional specificity encoded in sequences? How is sequence space explored by protein evolution? How does novel protein function emerge in the course of evolution?

Over the last years, the accumulation of massive sequence data thanks to next-generation sequencing together with advances in data-driven modeling using statistical and machine learning and artificial intelligence, have started to open radically new avenues in this field. While some questions seem to have obtained very advanced solutions – e.g. protein structure prediction from sequence using AlphaFold – finer functional and evolutionary questions remain highly challenging. Our project, to be carried on in a close collaboration with the Tokuriki lab at the University of British Columbia in Vancouver, Canada, who is developing world-class experimental techniques for high-throughput protein synthesis, assaying and experimental evolution, is based on joint modeling and experimental efforts, to

1. **resolve the fine organization of sequence space** of protein families using advanced data-driven modeling techniques inspired by statistical physics and artificial intelligence;
2. construct quantitative evolutionary models to **explain how novel protein function emerges** via the stochastic exploration of sequence space;
3. design methods for **coevolution-aware ancestral sequence reconstruction** (ASR) to overcome the limits of current ASR approaches assuming independent-site evolution.

Our modeling efforts are based on **generative statistical models** $P(a_1,...,a_L)$ for protein sequences $(a_1,...,a_L)$ of aligned length L belonging to one protein family. Informally speaking, a model is generative if artificial sequences sampled from $P(a_1,...,a_L)$ are statistically hardly distinguishable from the natural sequences in a protein family. The negative log-probability $E(a_1,...,a_L) = -\log P(a_1,...,a_L)$ can be interpreted as a **sequence landscape**, which assigns low values of E to biologically functional sequences, and high E to dysfunctional sequences. Using DCA as a generative model, M.Weigt and coworkers have recently shown that generated sequences are, with considerable probability, active *in vivo*, while simpler profile models fail to generate functional sequences[1].

Protein evolution proceeds via an interplay between **random mutations at the genetic level, and selection at the protein level**. We can therefore model evolution as a **stochastic process in a sequence landscape**. We have shown that this idea leads to evolutionary models, which thanks to the inclusion of **epistasis** outperform qualitatively and quantitatively existing independent-site evolution models[2]. Our goal is to use Monte Carlo sampling and **transition path sampling** in our sequence landscapes, in order to find connecting paths in sequence space changing protein specificity, which will be characterized experimentally. We also want to use similar techniques to **reconstruct evolutionary paths from ancestral to extant amino-acid sequences**.

This will allow us to address some of the most fundamental and fascinating questions in evolutionary biology: (*i*) quantify the role of epistasis in protein evolution, in particular the **contingency** of mutations to the sequence context shaped by prior mutations, and the **entrenchment** by subsequent mutations; (*ii*) investigate the **emergence of novel protein function** when the reconstructed ancestral and the extant protein have distinct functional specificities, and reconstruct potentially **promiscuous ancestral intermediates**.

[1] Russ et al. "An evolution-based model for designing chorismate mutase enzymes" *Science* (2020)
[2] Bisardi et al. "Modeling sequence-space exploration..." *Mol Biol Evol* (2022)

Logistics

We are seeking two highly motivated and passionate candidates, **for one PhD and one postdoctoral position starting in the fall of 2023**. The PhD could also begin with an internship in spring 2023. The selected candidates will work with us on the above project, and in close collaboration with the Tokuriki lab.

We are currently both located in Paris in nearby institutions (Sorbonne University, ENS Paris and ParisSanté Campus), and have been closely collaborating on this topic for a few years, jointly supervising the PhD of M. Bisardi and J. Trinquier and the postdocs of A. Muntoni and S. Cotogno. In this configuration, the logistics is very simple as the candidates would spend part of their time in both institutions.

However, **an important point** to be noted is that it is possible (although still far from certain) that F. Zamponi will move to the "Sapienza" University of Rome in the fall of 2023. **Should this happen**, the proposed logistics is the following:

- The PhD can optionally begin with an internship in spring 2023 while we are still both in Paris. The funding for the PhD can then come, depending on the candidate and the opportunities, from several sources: an EDPIF or EDITE contract in Paris, or a contract from Sapienza in Rome. Should all these options fail, we also have funding to finance the contract on our own grants. In all cases, we would activate an **International Dual Degree PhD** ("cotutelle" in French or "cotutela" in Italian). The PhD student would then spend half of their time in Paris and the other half in Rome, with a flexible schedule.
- The postdoc will be funded by a two-years contract located at ENS Paris and managed by CNRS. As such, the postdoc would be administratively based in Paris, but we expect them to travel frequently to Rome and if possible spend extended periods of time there (from a few months up to a full year).

It is important that candidates carefully consider these logistics and are open to spend part of their time in Paris and part in Rome.

How to apply

Please send us your application by email by **November 30, 2022**, including your CV, a few lines of motivations, a very concise research summary (for the postdoc), and a list of people that can be contacted for references (letters do not need to be sent at this stage). Please specify if you are interested in the PhD or postdoctoral position.

We will pre-select a few candidates that we will invite for an interview in December or January, and at this stage we will ask for recommendation letters.

We are strongly committed to supporting diversity and inclusion in early careers in research. Applications from young candidates (for a first postdoc after the PhD), women, and underrepresented minorities are thus especially welcome.