Unsupervised inference models for structural and functional properties of protein sequences

Synthesis

Ph.D. candidate: Matteo De Leonardis
Supervisor: Prof. Andrea Pagnani

February 25, 2025

Proteins are fundamental macromolecules that perform specific tasks inside living organisms. Their composition is highly specific to the function and it is encoded in the genome of each individual. Each protein is structured as a chain of simpler molecules called amino acids. There exist 20 possible amino acids, each with different chemo-physical properties. The complex interaction network between amino acids with different properties determines the shape that the protein assumes which corresponds exactly with the 3D structure they need to interact with other molecules in the space and carry out their function. Understanding how the composition affects the 3D structure is a long-standing goal in Computational Biology which has implications for several biomedical technological applications and would give precious insights on how genotype and phenotype expression are connected in living organisms.

Novel experimental techniques, like high-throughput sequencing, made it possible to screen large libraries of sequences boosting the production of high-quality data from such screening experiments that can provide information about the activity of sequences that are tested for a specific phenotypic trait. Despite the remarkable technological improvements experienced in the last years, the number of sequences that can be tested in these experiments remains still a tiny fraction of the set of all possible sequences. Therefore, the rise of sequencing data produced by screening experiments demands efficient and robust Machine Learning (ML) models that maximize the amount of information that can be extracted by analyzing them.

Two categories of experimental techniques are usually employed in screening experiments: Deep Mutational Scanning (DMS) and Directed Evolution (DE) experiments. In the former, a selection strategy is implemented that discards all sequences that don't exhibit the desired phenotypic trait. This selection strategy is iteratively applied to a pre-determined library of sequences. Sequencing the library during this procedure allows us to assess how its composition changes due to the selective pressure and extrapolate information on the activity of the sequences. DE experiments follow a similar pattern, but instead alternate the application of the selection strategy to the introduction of random mutations

along the sequences, in such a way to mimic real Darwinian evolution. These experiments usually start from a single sequence, and the trajectories followed by the experimental evolutional process spontaneously create a library of tested sequences.

Even if these experimental processes share several similarities with natural evolution, they are intrinsically different. While established inference methods are available for databases of natural sequences, inference from screening experiments remains an open problem that lacks a general framework. The main issue is that laboratory evolution occurs on time scales that are incommensurably shorter than natural evolution. This introduces correlations between experimental sequences that are usually negligible for natural sequences but, in the case of these experiments, require the knowledge of the phylogenetic tree to accurately describe the evolutionary process.

The problem of finding an optimal inference model for laboratory evolution experiments has been intensively addressed in the last years, and the most common approach boils down to the adaptation of established models for natural sequences to describe also laboratory evolution experiments. Lately, some works introduced the idea of reframing the inference problem as a time series analysis. The sequencing data provide temporal snapshots of a system evolving according to some unknown dynamics. The difficulty of this task is reflected in the fact that time-series analysis is in general a harder task in any ML application. These studies explored the idea of modeling the evolutionary process as a Markov Chian where the sequence represents the state of the system at each experimental round, and what we can observe as experimental data is the empirical frequency of the states at each time.

In this thesis, several unsupervised methods for inference are proposed and analyzed. Some of them are completely new, while one of them follows from previous work and has been adapted to deal with varying-length sequences. This model has been applied to a DMS experiment performed on variants of Adeno-Associated Viruses with different capsids to test their *viability*, the capacity of the capsid to remain stable for the packaging of the genome inside it. The model describes a generic realization of a generic DMS experiment and assumes that each sequence behaves like a two-state system and the probability of finding it in the *viable* state p_s follows the Fermi-Dirac statistics:

$$p_s = \frac{1}{1 + e^{E_s - \mu}} \tag{1}$$

This working hypothesis allows the model to be completely generic about the fact that p_s is regarded as a binary or a continuous variable. In fact, depending on the "temperature of the data", the model sets in one or the other case. The experiments allow for point-wise mutations in a specific capsid region of 28 amino acids but also allow for single insertions between any of these sites. This creates variants of different lengths, and our model must accommodate this feature. We use a Convolutional Neural Network (CNN) model to express the dependence of the *viability* on the variant sequence. We test our model in predicting the *viability* of sequences that were used for training, and we

compare our results with the case in which the same CNN is used in a standard ML approach: as a supervised method for a regression problem.

Except for this one, the other models are designed for DE experiments and are evaluated on two different datasets collected to test important enzymes for the activity of cells. One molecule is Dihydrofolate reductase (DHFR) which which performs an important function for cell metabolism: it is involved in the reduction of dihydrofolic acid into tetrahydrofolic acid via NADPH. The other one is Beta-Lactamase PSE1, which provides bacteria with resistance to a class of antibiotics called *beta-lactam*. Both experiments test antibiotic resistance of variants of these two enzymes in E. Coli. in the presence of low concentrations of antibiotics. A random mutagenesis technique called *error-prone PCR* is employed in both cases to introduce random mutations during the experiments.

In a first work, the evolutionary process is modeled as a sequence of steps of *Generalized Glauber dynamics* for a system of q-spins with pairwise interaction. We express the probability that at site i of the sequence, the mutation $a_i^t \to a_i^{t+1}$ takes place once the remaining context of the sequence is fixed

$$P_i(a_i^{t+1}|\mathbf{a}^t) = \frac{1}{Z_i(\mathbf{a}^t)} \exp\left[h_i(a_i^{t+1}) + \sum_{j\neq i} J_{ij}(a_i^{t+1}, a_j^t) + \mu\delta(a_i^t, a_j^t)\right] \qquad (2)$$

Unlike standard Glauber dynamics for a Potts model, we added the term $\mu\delta(a_i^t, a_j^t)$ that, for $\mu > 0$ takes into account phylogenetic correlations and keeps them separate from the fitness contribution due to interactions of sites of the sequence. We generalize the model to the case in which we have missing data in the time series and we compare the results of contact predictions with other similar inference models.

A second method is presented and analyzed. It combines dynamical modeling with a latent space representation of the sequences to retain only those features that are relevant to the protein's activity. The realization of the experiment is described as a diffusion process involving a random walker in the presence of friction, this model is commonly known as *Ornstein-Uhlenbeck process*. This framework aims to describe the interplay of noise and the external force that drives the particle in analogy to what happens for evolving sequences due to the interplay between random mutations and selective pressure. One of the most interesting features of this method is the ability to integrate the information from a database of natural sequences with the experimental ones. In fact, to choose how sequences are encoded in the latent space, we perform Principal Component Analysis (PCA) on natural sequences to select those patterns that are most relevant for fitness and we use them to project the experimental variants in a low-dimensional space. As long as the sequences are mapped in this space through a linear mapping, it is possible to find an approximate relation between the parameters of an equilibrium Gaussian distribution in the latent space and an equilibrium Potts model distribution in the sequence space, and then try to predict contacts in the 3D structure of the protein. As a validation, we show how, compared to other inference methods that don't use the

information from natural sequences, our model successfully integrates this information and remarkably improves the prediction ability with respect as one would expect. In addition, our model detects some contacts that are not found by standard methods that use only natural sequences. In other words, can extract some piece of information that remains hidden from other methods that separately use natural or experimental sequences for contact prediction.

In the last part of this thesis, a novel method for Ancestral Sequence reconstruction (ASR) is presented. Commonly, for the sake of computational efficiency, sites are assumed to evolve independently in most of the well-established approaches. Our model, instead, releases this assumption and uses an auto-regressive parametrization to preserve the computational efficiency. We use an approximated dynamic to parametrize the dynamics of a single site once the context of the sequence is fixed. We implement a modification of Felsenstein's pruning algorithm for time-irreversible processes, to reconstruct the state of any internal node of a phylogenetic tree, with a Maximum A-Posteriori estimate, one site at the time exploiting iteratively the auto-regressive parametrization. We apply this method to the DE experiment that tests PSE1. We compare how good this model is in reconstructing the reference sequence of the experiment in a simple setting, compared to a state-of-the-art method for ASR. We show that our model can be more accurate in reconstructing some mutations that are more difficult to predict where, arguably, the evolutionary constraints arising from the interactions with other sites are not negligible.