POLITECNICO
DI TORINO

Modeling the stochastic dynamics of protein evolution experiments using protein sequence landscapes

Master Degree in Physics of Complex Systems

Supervisors

Prof. Martin WEIGT

Prof. Francesco ZAMPONI

Prof. Andrea PAGNANI

Candidate

Leonardo DI BARI

July 2023

Table of Contents

Chapter 1

Introduction

Proteins are fundamental macro-molecules that are involved in a variety of vital functions in living organisms. They primarily consist of a linear amino-acid sequence, which allows the molecule to fold into a 3D structure and perform its function thanks to its chemical and physical properties. In this work, we are interested in understanding the sequence statistics and its evolution over different timescales. An interplay of mutations and selection shapes the amino-acid variety over the course of history. Understanding the stochastic dynamics of protein evolution is essential to the comprehension of the diversification of life and the emergence of new protein functions. Recently, the use of data-driven fitness landscapes and statistical physics methods to create a quantitative theory of protein evolution has gained more and more importance, leading to promising results [1, 2, 3].
Our aim is to numerically simulate protein evolution and compare the results with different experimental [4, 5] and natural data features, such as Hamming distance, contact prediction, and several orders of correlation statistics.
A Markov chain is used to describe the mutational dynamics, and a "sequence landscape" constructed from naturally occurring sequence variants is used to model the selection. The simulations start with a specific wild type protein and new proteins are designed using previously learned parameters to reproduce different stages of natural and in vitro evolution, which include experimental evolution at different rounds (10% of sequence mutations), and natural evolution observed in homologous sequences ($70 - 80\%$ of sequence variation).
After having confirmed the validity of our generative model both locally and globally, we search for emergence of epistatic signals at intermediate scales trying to give an intuition of the important timescales that rule the dynamics.
In conclusion, this project aims to provide relevant insights on the stochastic dynamics of protein evolution, which is essential to understanding the diversity of life.

1.1 A brief reminder on essential biology concepts

The following sections shortly introduce some key topics in biology, preparing the path for the core of our analysis.

Proteins

Proteins are large biomolecules composed of amino acids; they play a crucial role in biological processes such as catalysis, regulation, and structure.
Each protein can be depicted as a unidimensional polymer chain of amino-acids that folds into a three-dimensional convoluted structure (known as "tertiary structure") by varying the distances of different amino-acid sites. Such structure is closely correlated with the function performed in the cell, hence it is necessary to deepen our knowledge on the former to get more information on the latter. This connection is what made researchers suppose long time ago that the structure solely depends on the small building blocks composing the chain [6].
We can represent such relation in a simple way by saying that the information on the folding of a protein is hidden by a single phrase written in a 20-letter alphabet: the amino-acid sequence.

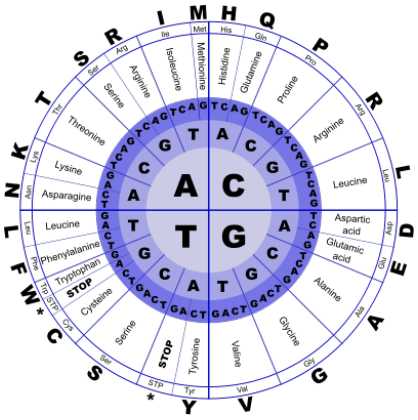

Figure 1.1: A schematic wheel representation of the genetic code: the three inner circles represent first, second and third nucleotide in the codon triplet.

Genetic code

What is the relation between protein and DNA? To answer this question we must talk about the set of rules by which the information encoded in DNA is translated

into the sequence of amino acids that make up a protein: the genetic code. The relationship between nucleotides and amino acids in this code is that specific codons, made up of three nucleotides, correspond to specific amino acids. In the standard genetic code, there are 64 possible codons, but they code for only 20 amino-acid as we can see in Fig. 1.1. Hence it is common to say that genetic code is "redundant": different codons can lead to the same amino-acid. The codons are read starting from a fixed point on the DNA molecule called the start codon and continuing until a stop codon is encountered.

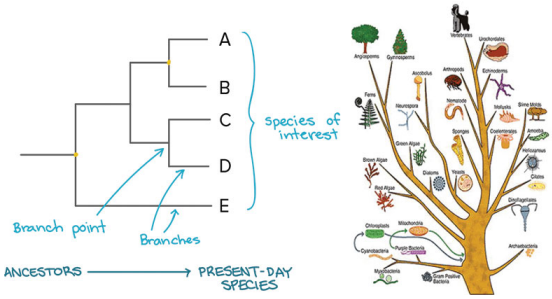

Figure 1.2: Pictorial representation of a phylogenetic tree.

Phylogenetics and protein families

Phylogenetics is a key field that uses inference methods, including traits like morphology, DNA, and protein sequences, to reconstruct the evolutionary relationships between species over time [7]. The resulting representation is known as a phylogenetic tree (Fig. 1.2), where nodes represent species or hypothetical ancestors, and branches represent genetic connections. Proteins also undergo mutation and can be organized into phylogenetic trees, referred to as protein families. Sequences within the same family are termed "homologs" and are believed to share a common evolutionary ancestry. The growing availability of labeled protein databases, driven by advances in sequencing techniques, provides a wealth of information for data-driven modeling and analysis.

Natural Protein Evolution

Biological evolution is a complex process which is yet to be completely understood. However, from an effective point of view it can be efficiently described as a highly stochastic process in a fitness landscape acting on populations. Mutations in DNA lead to changes in protein sequences, which, along with selective pressures,

determine the survival and reproduction of organisms. Most mutations decrease fitness, while some others are beneficial and confer a fitness advantage when under selective pressure. In an equilibrium between the two, protein evolution can be viewed effectively as a neutral walk on the fitness landscape. Proteins serve as an excellent proxy for studying evolution due to their robustness to mutations and sensitivity to even slight amino acid substitutions for what regards their functionality. Furthermore, proteins exhibit slow changes over time and their structure are highly conserved across species, making them valuable for studying evolutionary relationships.

Experimental Protein Evolution

To carry on our review on evolution, we now ask if it is possible not only to look at long time scales by collecting data from different species, but also to try to reproduce the stochastic dynamics of evolution at shorter time frames. One powerful approach is directed evolution, which involves introducing mutations into a protein sequence and then selecting for variants that exhibit a desired function or property. Experiments involve several steps, beginning with the creation of a library of variant proteins. This library is usually generated by introducing mutations into the gene encoding a target protein, either randomly or at specific positions using techniques like error-prone PCR or site-directed mutagenesis. The resulting variants are then screened or selected for the desired function or property.
Screening involves testing each variant individually for the desired function or property, while selection involves subjecting the entire library to selective pressure (e.g. growth in a specific environment) and isolating the variants that exhibit the desired phenotype. Variants that pass the screening or selection step are then characterized further to understand the mechanisms underlying their improved performance.
This process of screening, selection, and characterization is typically repeated over many rounds, with each one involving further mutation and selection of the most promising variants.

Epistasis

When talking about mutational effects on protein sequences, it is necessary to discuss about epistasis. Briefly speaking, this concepts refers to mutations involving interactions between different protein sites. Contrary to what one normally should expect, the outcomes of mutations are not additive. Epistasis is related to a non-linearity in the mapping from sequence to biological property and it can restrict the trajectories available to an evolving protein or open new paths to sequences

and functions that would otherwise have been inaccessible. We can talk about negative or positive epistasis whether the effect of a pair of mutations results in a larger or smaller phenotype given by the sum of the two single mutations. Actually, characterizing such phenomena is a difficult task and its precise role in protein evolution is hardly understood. This work aims to contribute to a quantitative theory of epistasis, adding to different studies such as [8, 9, 10].

Betalactamase family

Before going on with the description of protein evolution experiments, it is necessary to introduce the protein family that has been taken under our lens and show why it is crucial to understand its characteristics.
The beta-lactamase family is a group of enzymes that are commonly found in bacteria and are responsible for conferring resistance to β-lactam antibiotics, such as penicillin and cephalosporins. Antibiotics all have a common element in their molecular structure: a four-atom ring known as a beta-lactam (β-lactam) ring. Through hydrolysis, the enzyme lactamase breaks the β-lactam ring open, deactivating the molecule's antibacterial properties.
In the following experiments [5, 4] two particular proteins belonging to this family were used: TEM-1 and PSE-1. They are two prominent members of this family that have been extensively studied. TEM-1 was the first beta-lactamase enzyme to be characterized and is still one of the most widely distributed beta-lactamase enzymes in bacterial pathogens. On the other hand, PSE-1 is a broad-spectrum beta-lactamase that is capable of hydrolyzing a wide range of beta-lactam antibiotics.
Studying beta-lactamase enzymes is crucial for combating antibiotic resistance, developing new antibiotics, and understanding mechanisms of resistance in bacterial pathogens, thereby safeguarding public health.

Chapter 2

Data and Methods

In the following chapter we will deal with the data-sets utilized for our study and we will address Direct Coupling Analysis, a powerful statistical-physics inspired tool used to infer precious information regarding protein and DNA evolution. Such method aims at learning from the large protein databases a probability distribution of the sequence in a Boltzmann-like fashion. This probabilistic model assigns an energy landscape that characterizes how probable is to observe each sequence; it can provide further insights on biological features such as residue-residue contact prediction or the effect of mutations on fitness landscape.
Furthermore, the same distribution can be subsequently used as a generative model for artificially sampling proteins that are statistically indistiguishable from the ones contained in the training set.
Our new contribution will be to provide this static framework with a new mutational dynamics simulating the evolutionary history of sequences. This process relies on deep biological concepts which ensure its suitability in mimicking complex phenomena undergoing at the nucleotide level, while still asymptotically reaching equilibrium and recovering characteristics of the training set. As an example, mutation, insertion and deletion mechanisms should be considered at the level of DNA, taking into account the genetic code.

2.1 Experimental datasets

Stiffler Experiment

We discuss an experiment [4] realized by Stiffler et al. in 2020. They concentrated on two proteins, Betalactamase PSE-1 and aminoglycoside acetyltransferase AAC6, that share the antibiotic resistant feature. We will just focus on the first one because we want to keep our analysis confined to the Betalactamase family.

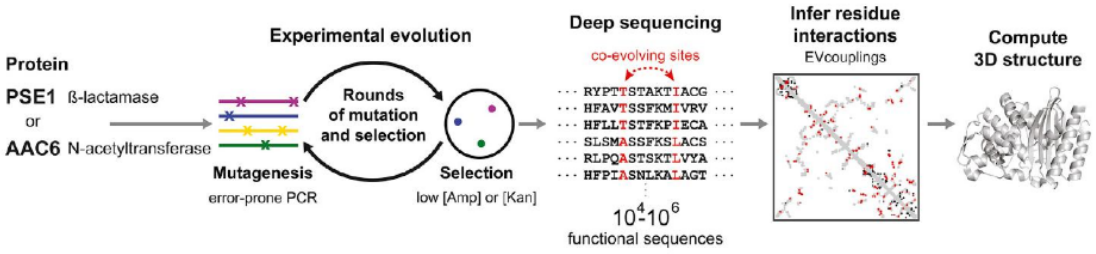

Figure 2.1: Schematic representation of Stiffler experiment taken by [4].

A pictorial view of the workflow of the experiment is shown in Fig.2.1. In order to facilitate differentiation, error-prone PCR was used with a high-mutation rate followed by selection of functionality through external pressure realized by $6mg/mL$ ampicillin. This dose corresponds to low selection: not only the best proteins survive. Therefore, the experiment explores the neutral space, but it does not optimize for resistance. The process was iterated 20 times by taking the survived proteins of a round as the input sequences for diversification in the next one. The sequencing was performed at $10th$ and $20th$ round resulting in two diversified MSAs. These latter ones were then used to learn a Potts model leading to satisfactory contact prediction results.
Stiffler's experiment managed to gather significantly more sequences ($\simeq 465000$). A small caveat must be made on this number: it was initially reduced down to $\simeq 165000$ because of a strong cleaning performed by Stiffler et al., but we choose to use the complete data-set (with some standard pre-processing) for our analysis.

Fantini Experiment

A very similar experiment has been carried out by Fantini et al. [5]. After processing the data, the researchers were able to create a highly diversified library of 10^5 sequences at a distance of nearly 10% from the TEM-1 wildtype belonging to Betalactamase family. Finally, the data was used for the main scope of this paper, which was to predict residue-residue contacts. The results were significantly worse with respect to those of Stiffler.

2.2 Direct Coupling Analysis

DCA is based on two main evidences regarding a single protein family: all sequences share the same 3-D structure and they have high intra-family variability at the level of amino-acids. The similar function operated by these proteins in the cell suggests that during evolution some common features have been preserved. Such

characteristics might be identified as constraints under which the folding in a three-dimensional structure takes place. It is precisely these last ones that are usually statistically inferred by exploiting the huge amount of sequences and intra-data variability that can be found in protein databases. More precisely, constraints

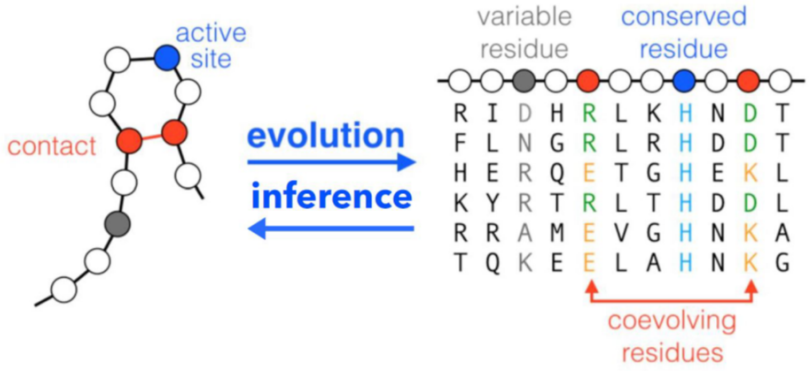

Figure 2.2: Schematic illustration of the relation between three-dimensional structure and sequence variability in protein evolution, taken by [3].

can be both local (linked to functionality and binding) or global (associated to stability and folding). Along evolution, we might have amino-acid specific residues that never mutate (i.e. the blue-dotted site in Fig. 2.2). Nevertheless, we can also have couples of amino-acids that variate accordingly (i.e. the red-dotted residue in Fig. 2.2). This can be easily addressed as a co-evolutionary effect related to residue-residue contacts: a pair of amino-acids in contact in the folded state must have correlated mutations over evolutionary timescales in order to maintain the necessary stability [11, 12].

Multiple Sequence Alignments

The starting point of DCA is the building of a good data-set on which we perform our inverse statistical physics procedure. The approach is to take the protein family from the database and organize it into a matrix of $M \times N$ dimension in which each row is a protein and each column is a residue site. We will then depict each element/amino-acid of the MSA as a_i^m (with $m = 1, ..., M$ and $i = 1, ..., N$) taking 21 possible symbols that represents the typical letters of the amino-acids plus a gap $'-'$. As a matter of fact, during evolution events such as insertion or deletion of single sites cause the alignments to have some holes or extra tails. In order to take account of this, tools like the hmmsearch command from the HMMer software suite build the MSA directly from the protein database by using scores of most probable

alignments [13]. This task is not easy at all and we might see some references to these procedure in the following chapters.
To get acquainted with the scales we are exploring, we can safely consider the range of M to be up to 10^6 and the number of sites N from 50 to 500 in the case of PFAM domain family [14]. The amino-acid letters can be respectively encoded as integer numbers or as "one-hot" sequences where each letter is mapped into a unit 20-dimensional sparse vector of all 0's and just a 1 (with the gap being a null vector). The first encoding is more practical, whereas the second might be useful in cases in which we want all amino-acids to be equally distant like in Principal Component Analysis.
Observing this data-set, we can notice statistical features among columns and rows. The simplest empirical measures we can look at are:

- single-site frequencies $f_i(a) = \frac{1}{M} \sum_m \delta_{a,a_i^m}$
- double-site frequencies $f_{ij}(b,c) = \frac{1}{M} \sum_m \delta_{b,a_i^m} \delta_{c,a_j^m}$

The two quantities can be easily understood in terms of biology. One-point frequencies $f_i(a)$ depict the process of conservation in evolution. If a specific site plays a crucial role in functionality, such as binding to a particular substrate or having specific biochemical properties, then a mutation affecting this site would likely have negative consequences. As a result, that site would be conserved within the family. On the other hand, two-point frequencies $f_{ij}(b,c)$, when $f_{ij}(b,c) \neq f_i(b) f_j(c)$ demonstrate the phenomenon of coevolution. If two sites are in contact in the folded state of the molecule, then changes in one site due to a mutation may require changes in the other site in order to preserve the ability to form the bond. It is common to exploit highly amino-acid dependent sites to build correctly aligned MSA of proteins belonging to the same family. Their strong conservation suggests that they might have an essential role in making the protein fold or perform its function in the correct way.

Statistical Modelling

As we underlined in the previous paragraph, coevolution is evident through correlated occurrences of amino acids in different positions within a protein. Utilizing these correlations to predict the protein's three-dimensional structure and contact map has been a longstanding approach [15, 16]. However, it is challenging because correlations and mutual information in a multiple-sequence alignment do not directly represent amino acids in direct contact. Indirect correlations can arise when pair of positions are in contact with others, complicating the analysis.

The key assumption of this inverse statistical physics problem is to admit that proteins of same family (i.e. the rows of the MSA) can be considered as (not

necessarily i.i.d.) samples from a probability distribution [17, 3, 18] that takes the following Boltzmann-like expression:

$$P(a_1, ..., a_N) = \frac{e^{-H(a_1,...,a_N)}}{Z} \tag{2.1}$$

where N is the length of the sequence, the inverse temperature β has been absorbed in the Hamiltonian $H(a_1, ..., a_N)$ and $Z = \sum_{a_1,...,a_N} e^{-H(a_1,...,a_N)}$ is the partition function commonly encountered in statistical physics. The interpretation of the "energy" $H(a_1, ..., a_N)$ is quite easy: sequences with a lower energy will have a higher probability, hence they will be more frequently sampled, resulting in a statistical predominance in the alignment. In this sense, the Hamiltonian provides a high-dimensional landscape in which rarely observed sequences correspond to peaks, whereas functional frequently selected sequences populate the valleys.
One could argue whether this assumption is robust and acceptable, considering the complexity of the subject of our study. Regarding this, we refer to the brilliant work of Jaynes [19] where he explains how statistical mechanics approaches can be recovered starting from the information theory-based Maximum Entropy Principle. Indeed Jaynes showed that if we want to infer probability distributions that match specific empirical observables of a data-set, then the Boltzmann distribution is the least-biased one (i.e. the one with highest entropy) among them.

Potts model

The model we will consider when building an energy function for our sequences is the Potts Model:

$$H(a_1, ..., a_N) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j) \tag{2.2}$$

.It consists of "local fields" $h_i(a_i)$ modelling the amino-acid conservation at specific sites and "couplings" $J_{ij}(a_i, a_j)$ which should account for the description of co-evolutionary residues. The latter term is of fundamental importance when trying to predict residue-residue contacts. The inference of couplings is a complex computational problem as it involves calculating thermodynamic averages, which have to be in line with empirical values, from an exponentially large sequence space in a disordered model without any prior symmetry. In the context of Potts models, the effect of mutations can be analyzed easily by computing the energy difference

$$\Delta E(a_i \to c) = H(a_1, ..., a_{i-1}, c, a_{i-1}, ..., a_N) - H(a_1, ..., a_{i-1}, a_i, a_{i-1}, ..., a_N). \tag{2.3}$$

A negative ΔE accounts for beneficial mutations, whereas a positive one might describe deleterious changes in the amino-acid sequence. It is noteworthy that the

use of pairwise Potts Hamiltonians has previously been explored in the context of protein design [20, 21] to understand the distribution of sequences that fold into a known structure.

The parameters of the model are inferred from a training data-set through Maximum Entropy Principle (MEP) [19]: we choose a set of observables $O_\alpha(a_1, ..., a_N)$ of the data-set that we want to reproduce with our probabilistic model, constrain the parameters to reproduce such observables and select those that maximize the entropy of the distribution. In our specific case we decide only to impose:

- one-point frequencies $f_i(a) =< \delta_{a,a_i} >_P$
- two-point frequencies $f_{ij}(a,b) =< \delta_{a,a_i}\delta_{b,a_j} >_P$

for every possible i, j, a, b. For a precise justification and analysis of the learning procedure, please check Appendix A.

2.3 Generating artificial sequences

One of the most powerful features of DCA framework is the possibility of using the learned model parameters to sample yet unknown functional sequences of amino-acid. This aspect is quite remarkable given that only a very small subset of the enormous sample space ($\simeq 20^L$, where L is the number of amino-acids) results into working proteins. It has been shown [22, 23, 24] that through Boltzmann sampling from the learned probability distribution, it is possible to obtain new functional proteins that are highly different in terms of Hamming distance from the known ones. It is noticeable that such a fine-tuned quality like functionality can be almost recovered just by fitting one and two point frequencies.
There are several procedures that can be used to sample our new sequences from the known distribution, the most used one is Markov Chain Monte Carlo, which iteratively generates samples from a distribution of interest.

Sampling sequences

Once the parameters are correctly estimated and regularized, we are provided with a Boltzmann probability characterizing each sequence with a score. This can be used to recover functional protein sequences from an i.i.d. equilibrium sample and predict mutational effects. In this sense it helps to understand how selection acts on proteins in natural evolution. One novel introduction we made with respect to previous works is the choice of a new Hamiltonian (and consequently of probability

distribution to be sampled)

$$H(B) = H[A(B)] + T \log \mathcal{B}(A(B)) = H[A(B)] + T \sum_{i=1}^{L} \log \mathcal{B}[a_i(B)] \tag{2.4}$$

where we call $\mathcal{B}(A) = |\{B : A(B) = A\}|$ the number of nucleotide sequences B that correspond to the same aminoacidic sequence A. Because each codon independently codes for an amino acid, we have

$$\mathcal{B}(A) = \prod_{i=1}^{L} \mathcal{B}(a_i) \ , \tag{2.5}$$

where $\mathcal{B}(a_i)$ is the number of codons that code for amino-acid a_i.

This takes into account the fact that different codons can code the same amino-acids and favors with a higher probability/lower energy the sequences that have a low degeneracy, i.e. that can be coded by just a few combinations of nucleotides. Instead, amino-acids that have a high genetic redundancy are more rarely sampled. In this way, probability is correctly distributed among all possible codons. One assumption is that we do not have codon biases: the energy is defined only on amino-acid, it is blind to synonymous codons that are translated into the same amino-acid.

Regarding the sampling procedure, a challenge arises in maintaining detailed balance, which ensures that at equilibrium each process is in equilibrium with its reverse process. A Markov chain satisfying detailed balance is called "reversible" Markov chain and it must converge to an equilibrium distribution. In other words, if our algorithm satisfies detailed balance, we can use it not only to sample short-term evolution experiments, but we can directly recover natural evolution by reaching equilibrium. This means that the same model can safely and correctly describe a local fitness landscape in the vicinity of TEM-1 and PSE-1, but also reproduce the enormous global variability of the entire betalactamase family.

However, what about the stochastic evolutionary dynamics? This theoretical framework must be completed with a suitable mutational process which both reflects known evolutionary biological phenomena and ensures that we are asymptotically reaching the prefixed stationary distribution learnt on the training set of natural sequences. Our major contribution in this work is to develop such dynamics. Firstly, we have already seen from the definition of the Hamiltonian that we managed to take into account evolution at the level of nucleotides through the genetic code without introducing codon biases. In addition, we should introduce in our stochastic dynamics two other evolutionary phenomena:

- single-nucleotide mutations → one nucleotide is replaced by another
- deletions → an entire codon is replaced by a gap (3 nucleotides get deleted)

- insertions → an entire codon is replaced by a gap (3 nucleotides get inserted)

where gaps are positions in a protein sequence where one amino acid is missing. They can occur due to various reasons, such as insertions or deletions in the DNA coding sequence, or during the process of sequence alignment.
We see that one process happens at the level of a single nucleotide, whereas another acts on triplets, making the satisfiability of detailed balance (necessary to recover the stationary distribution) not trivially compatible with simple dynamic rules currently used in MCMC simulations.
Here we propose the following solution with a more detailed derivation presented in the appendix B.

We choose to work with Gibbs and Metropolis sampling, combined with Markov Chain Monte Carlo (MCMC). In Gibbs sampling, variables are sampled iteratively from their conditional distributions, considering the current values of other variables. This enables sampling from the joint distribution of all variables, even when obtaining the joint distribution analytically is not feasible or computationally expensive. On the other hand, Metropolis sampling consists of proposing changes in the studied configuration and accepting them through a precise score based on the effect of the move on the model energy. For our purposes, we develop a mixed sampler that proceeds in the following way:

- start with a protein sequence
- with probability p do a Metropolis move modelling insertion/deletion of a codon
- with probability $1 - p$ do a Gibbs move modelling single-nucleotide mutations
- iterate the process with the updated sequence as input.

The Gibbs move proceeds in the following way:

- Choose a nucleotide position i_k at random (considering non-gapped positions only)
- Compute the probability of all possible mutations in that position $P(b'_{i_k}|B_{-i_k})$
- Sample the new nucleotide b'_{i_k} according to the probability density $P(b'_{i_k}|B_{-i_k})$

where $P(b_{i_k}|B_{-i_k})$ is the probability of a single nucleotide mutation, conditional to all other nucleotides.
Contrarily, the Metropolis move works on the codon space in the following way:

- Choose a codon $b = (b_1, b_2, b_3)$ at random
- Propose a move $b \to b'$ to a new state through a proposal matrix $\Omega(b \to b')$

- Accept the move according to the probability $p(b \to b')$

where $p(b \to b') = \min\left(1, e^{-\beta[\mathcal{H}(B')-\mathcal{H}(B)]}\right) = \min\left(1, \frac{\mathcal{B}(A(B))}{\mathcal{B}(A(B'))} e^{-\beta[H[A(B')]-H[A(B)]]}\right)$. The proposal matrix is the following

$$\Omega(b \to b') = \begin{cases} \alpha, & \text{if } b, b' = G \\ \beta, & \text{if } b = G, b' = C_i \text{ or } b = C_i, b' = G \\ \gamma, & \text{if } b = b' = C_i \\ 0, & \text{if } b = C_i \text{ and } b' = C_j \text{ with } i \neq j \end{cases}$$

where G is a gap and C_i is one among the possible 64-codons observed in nature. This matrix has been constructed in order to admit insertion and deletions, disallow substitutions of codons ($C_i \to C_j$), while still being normalized and symmetric. From an efficiency standpoint, one could argue that the non-null diagonal makes the algorithm quite slow since it proposes useless moves that don't change the sequences. Unfortunately α and γ are necessary for normalization, still they can be reduced by maximizing insertion/deletion proposals choosing $\beta = \frac{1}{64}$, which consequently imposes $\gamma = \frac{63}{64}$ and $\alpha = 0$.

Chapter 3

Results

3.1 Short term sampling

Here we want to show that our model using the previously described mixed sampling that respects detailed balance is able to reproduce the statistics of experiments on protein evolution with a sequence variability of around 10%. Indeed, it has already been shown recently by Bisardi et al. [25] that with a slightly different (non-satisfying detailed balance) sampling technique, it is possible to recover results of two important protein evolution experiments [5, 4].
Our aim is to use parameters learned on the beta-lactamase family of natural sequences and ensure that the new sampling procedure manages to predict the short-term evolution features of the two cited experiments. For the following analysis we will set $p = 0$ in the sampler, so using only Gibbs moves. This choice is motivated by the fact that gapped sequences have been cleaned away in the experimental data-sets.

Tuning hyperparameters

The first thing that we should take care of when trying to reproduce the statistics of in-vitro protein evolution is linking our model parameters to features of the experiment and correctly tune them. Here, we consider the sequencing rounds and the selective pressure previously described. We can associate the first with the amount of steps performed in our MCMC dynamics and the second with an effective temperature used for computing the Boltzmann probability acceptance in the simulation. Obviously, more steps imply a bigger distance from the original sequence and high temperature allows all mutations, even highly deleterious ones, to take place. Unfortunately the two hyperparameters of our dynamics are correlated and must be tuned together. As a result of this, we decide to perform a grid-search

in the parameter space using as error score the sum of squared relative errors on the mean Hamming distance and mean energy between the experimental and artificial set.

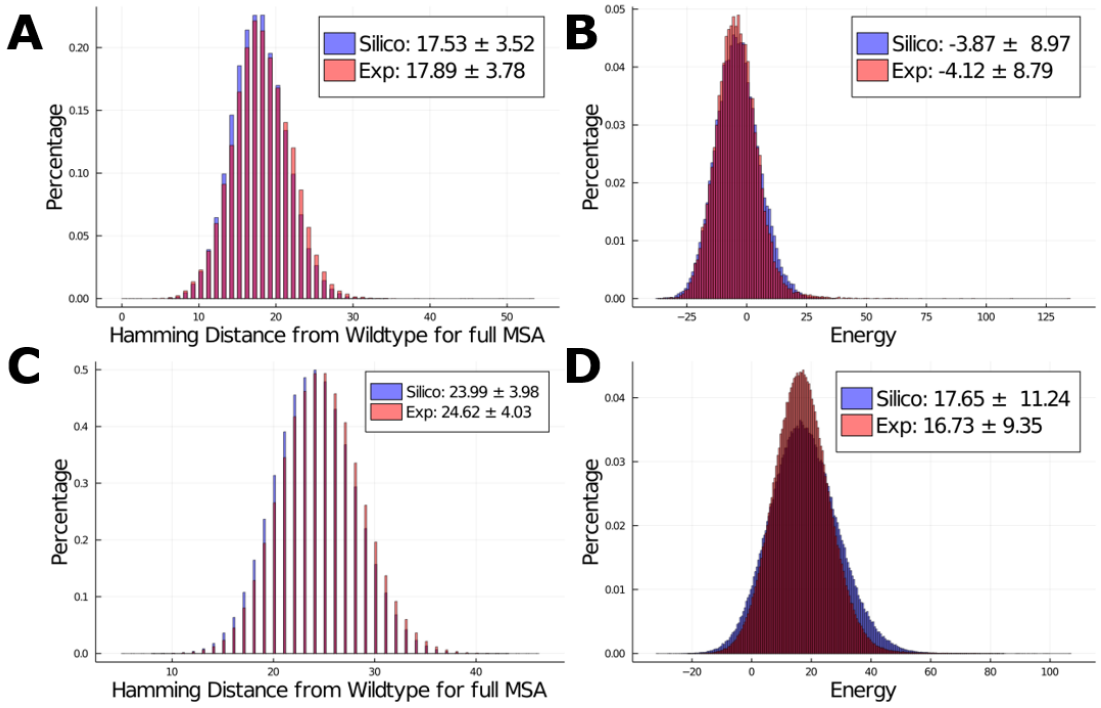

Figure 3.1: Comparison of statistical energies and Hamming distances from wildtype for in vitro and in silico alignments: (A) Artificial and experimental Hamming distance distribution for Stiffler (round 20), (B) Artificial and experimental energy distribution for Stiffler (round 20), (C) Artificial and experimental Hamming distance distribution for Fantini (round 12), (D) Artificial and experimental energy distribution for Fantini (round 12).

As we can see in Fig. 3.1 the distributions of statistical energies and Hamming distances from wildtype are quite accurately recovered by the artificial sequences. It is important to notice that the model fits extremely well the variance of such quantities even if the grid-search is performed only taking into account the first moments of the distributions.

Features comparison

Now that we have correctly linked the mutational artificial dynamics to the parameters of the experiment, we move on with our comparison between the in-vitro and in-silico data-sets.

One-point frequency

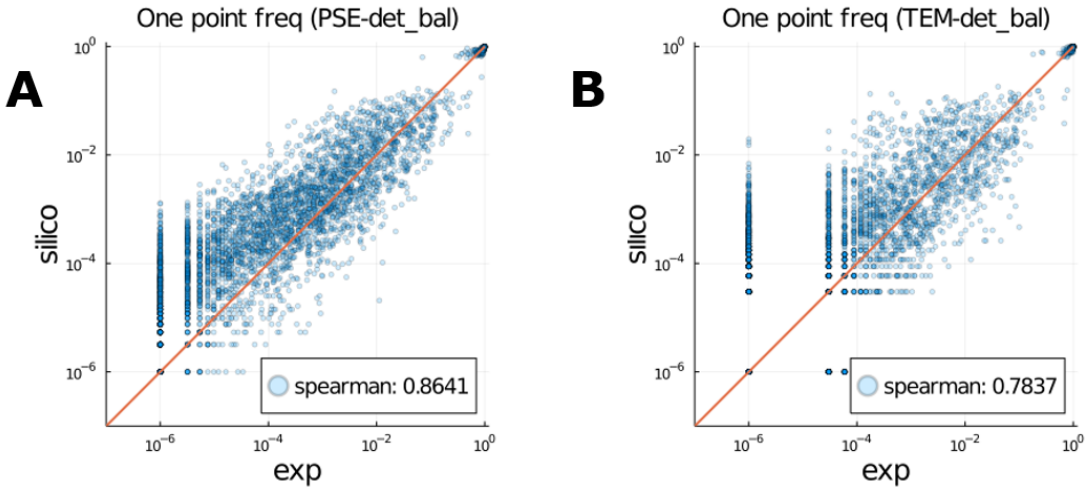

Figure 3.2: Comparison of first order statistics between in-vitro and in-silico alignments: (A) Scatter plot of artificial and experimental one-point frequencies for Stiffler (round 20), (B) Scatter plot of artificial and experimental one-point frequencies for Fantini (round 12).

Firstly, we look at 1-point frequencies. As it is evident in Fig. 3.2, we have an extremely high Spearman correlation considering that our Potts model was learned on distantly diverged sequences, while the simulation correctly reproduces sequences that are closer to the wildtype than to any other protein in the training data. This correlation results to be higher for PSE due to the size of the alignment being one order of magnitude bigger than that of TEM. As a matter of fact, exploring more extensively the sequence space allows to more easily sample even rare combinations of amino-acids.
We can dig in these plots a bit more, questioning if our sampling model is performing better than the one developed by Bisardi et al.[25] that does not employ detailed balance. We concentrate on the clusters on the top right corners of Fig. 3.2 which is filled by the most conserved amino-acids.

Despite not seeing a major difference in the case of Stiffler round 20, the increase in correlation for TEM is quite remarkable from Fig. 3.3. From a visual inspection it is quite clear that there is a lower bound for the artificial sequences from Bisardi et al., resulting into a rigid cut of the scatter plot. Fortunately the issue is just

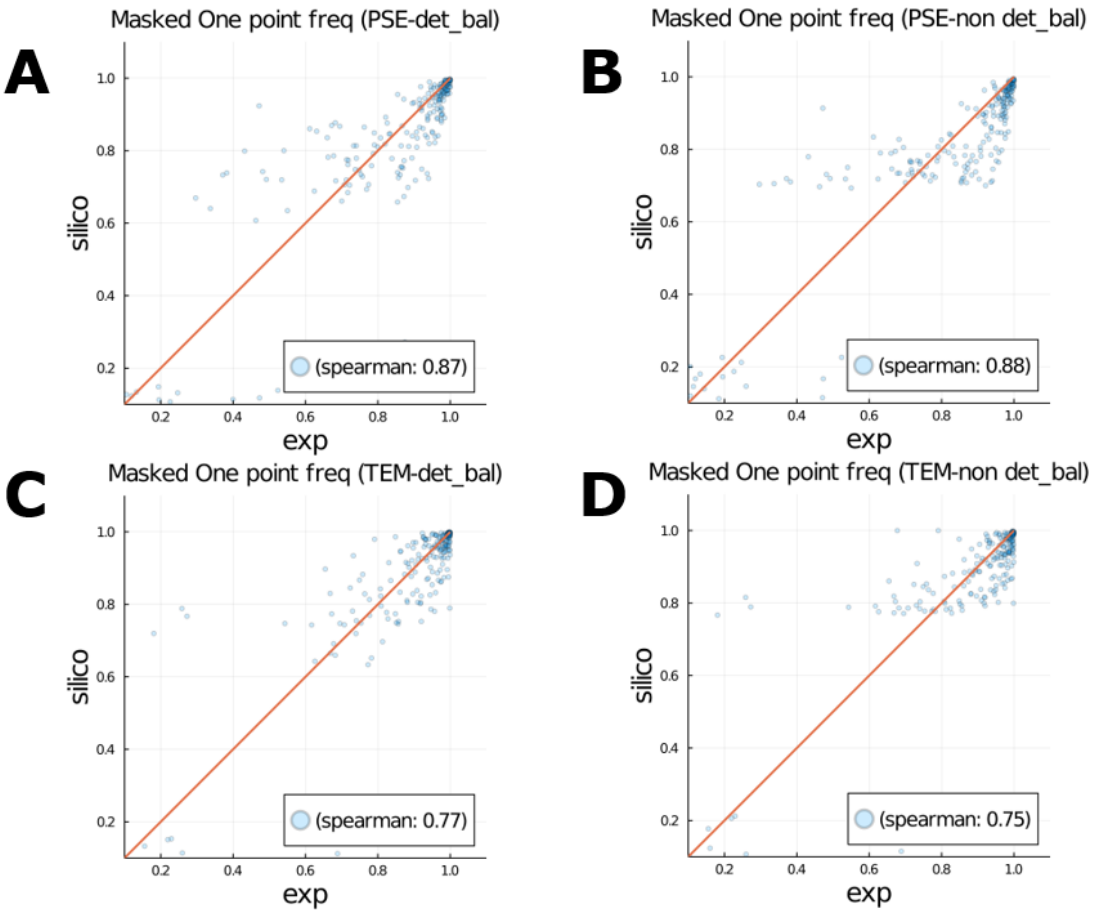

Figure 3.3: Comparison of first order statistics between in-vitro/in-silico alignments and detailed balance/non-detailed balance sampling technique: (A) Zoom of scatter plot of artificial (realized with detailed balance) and experimental one-point frequencies for Stiffler (round 20), by taking a subset of 10K proteins, (B) Scatter plot of artificial (realized from Bisardi et al.) and experimental one-point frequencies for Stiffler (round 20), by taking a subset of 10K proteins, (C) Scatter plot of artificial (realized with detailed balance) and experimental one-point frequencies for Fantini (round 12), (D) Scatter plot of artificial (realized from Bisardi et al.) and experimental one-point frequencies for Fantini (round 12).

related to intermediate frequencies. Instead, the new model appears to sample the site variability in a more appropriate way, with points closer to the ideal bisecting line.

Site entropy

A second feature we would like to explore is the site entropy. This is defined as

$$s_i = -\sum_{a=1}^{20} f_i(a) \ln(f_i(a)) \tag{3.1}$$

and it represents the variability of a position. As a matter of fact, e^{s_i} is the effective number of observed amino-acids in position i for a given multiple sequence alignment. To better display our comparison, we perform a scatter plot between artificial and experimental sequences. A particular distinction has been made regarding the sites: we call "non-mutated" the positions that share the same amino-acid between TEM and PSE wildtypes, whereas all the others are denominated as "mutated". Indeed, around 70% of sites are different between the two wildtypes, confirming the strong intra-family divergence of natural homologs.

From Fig. 3.4 we realize that entropy of non-mutated positions is definitely lower than the others as expected. In addition, it is notable that correlations are quite satisfactory ($\simeq 0.7/0.8$) given that we are using a model that has not been trained on the data related to the experiments. It is necessary to underline that we are correctly recovering the results of previous works [25] as the correlations are comparable. To conclude, we can say that we are able to model the local variability of protein residues, with a slight improvement with respect to previous analysis. In conclusion, we see that if the model hyperparameters for the evolutionary time (number of MCMC steps) and selection (artificial inverse temperature) are correctly tuned, the model reproduces fine statistical features of the experimental data, like the site specific probabilities of mutations, without ever having used the data in the learning procedure. This shows that, at least for short evolutionary times reached in the experiments, our model allows for quantitative predictions of evolution.

3.2 Long term sampling

In this section, we try to recover the main statistical features of natural evolution MSAs to further assess the validity and robustness of our generative model. We take as reference data-set the alignment of the PF13354 used for learning the Potts model parameters by Bisardi et al. [25]. In addition, we will use our mixed sampling technique with a $p = 1/2$ (perfect alternation of Gibbs mutation moves and Metropolis insertion/deletion ones).

Procedure

Our aim is to sample the global landscape that has been shaped by natural evolution by starting all our simulations in a local region. This is what proves that our

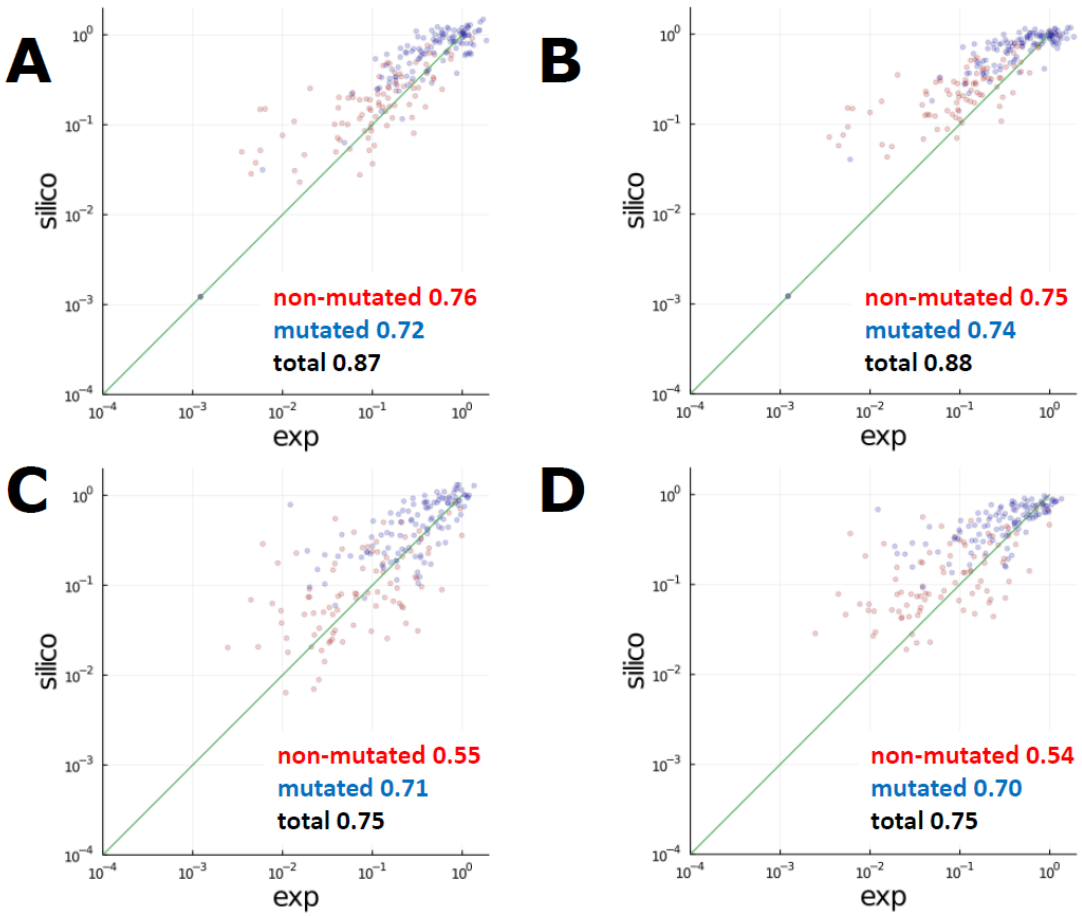

Figure 3.4: Comparison of site entropies of in-vitro/in-silico alignments and detailed balance/non-detailed balance sampling technique (the distinction between "non-mutated" and "mutated" sites has been done considering mutations between sequences of PSE-1 and TEM-1 wildtypes): (A) Site entropies of artificial (realized with detailed balance) and experimental proteins for Stiffler (round 20) , (B) Site entropies of artificial (realized from Bisardi et al.) and experimental proteins for Stiffler (round 20), (C) Site entropies of artificial (realized with detailed balance) and experimental proteins for Fantini (round 12), (D) Site entropies of artificial (realized from Bisardi et al.) and experimental proteins for Fantini (round 12).

model really can access information on the entire protein family irrespectively of the initialization of the sampling procedure.
As a matter of fact, we start our Markov chains in a single natural wildtype, take for example the PSE-1 betalactamase used in Stiffler experiment [4]. The sampling technique implemented in Appendix B respects detailed balance equations, hence it assures that after a certain amount of steps we start to sample the

correct equilibrium distribution, i.e. the natural sequences landscape. However the threshold after which we are correctly equilibrating is a-priori unknown. To reach the amino-acid variability of natural sequences we must extend the sampling procedure described in the previous section on much longer timescales, possibly raising some questions on the computational feasibility of the task.
To monitor our proximity to equilibration, we keep track of the dynamics realizing multiple sequence alignments at 112 different time steps chosen according to a logarithmic time grid and we compare their features with the natural evolution data-set. Despite being just a check on the evolution of the dynamics of the simulation, this data reveals to be densely rich in information.

Data-sets comparison

To monitor how well the sampled sequences are reproducing the reference data-set over time, we check different statistical features.

From Fig. 3.5 we can see that the simulated dynamics correctly reaches the variability of the natural datasets both for what regards the amino-acid frequencies and the pairwise Hamming distance distribution. In addition, the gap statistics is recovered with high correlation.

To assess the ergodicity of our sampling procedure, we show in Fig. 3.6 the principal component analysis of natural sequences and we project on it the last step of our simulations. In addition, we plot the trajectory of one sequence and highlight that it is correctly visiting the natural clusters. In conclusion, high statistical correlations saturating in time and ergodic exploration of the sequence space confirm the generative power of our model.

3.3 Emergence of epistasis at intermediate scales

After having validated our model against data from short and long-term evolution we can also explore all intermediate scales, where no experimental data are available. Before showing our results on epistatic signals, we must introduce a crucial definition that will be used in the analysis. Regarding epistasis, it is important to underline that the rejection/fixation of mutations at a specific site is strongly affected by the aminoacid present in the others. A good metric of how much a position is "bound" to its background is the context-dependent entropy

$$s_{i,CDE} = -\sum_{a=1}^{20} p_i(a|A_{-i}) \ln(p_i(a|A_{-i})) \tag{3.2}$$

that is basically the standard (context-independent) entropy where we substituted the simple frequencies with the probabilities $p_i(a|A_{-i})$ of aminoacids at site i given

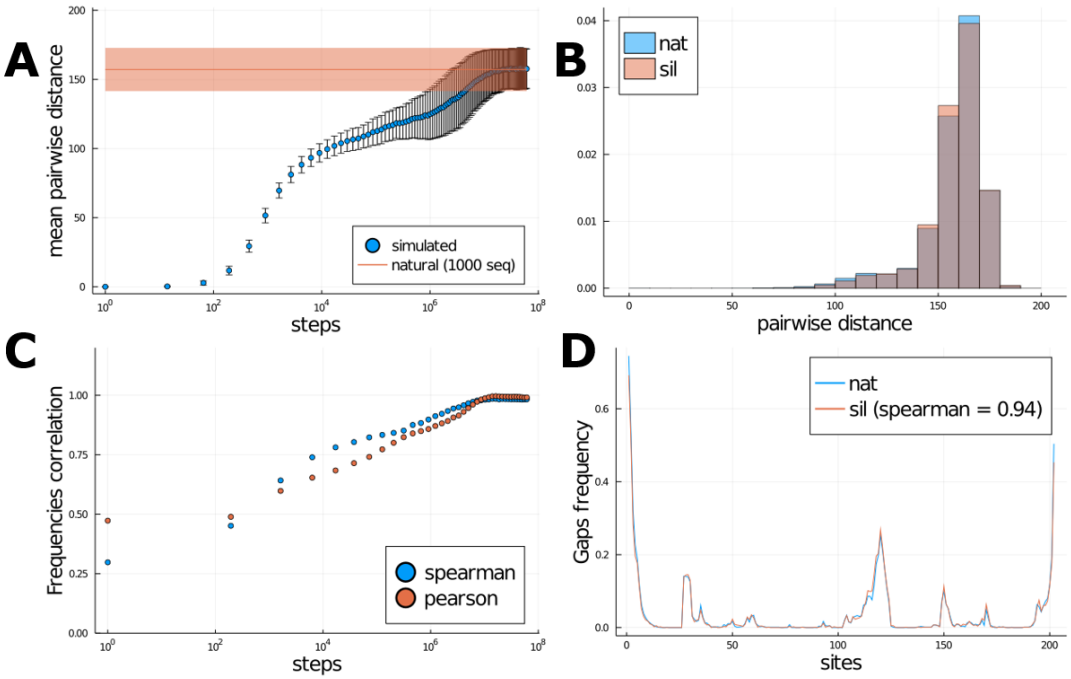

Figure 3.5: Comparison of statistical features for natural and simulated alignments: (A) Evolution of the mean pairwise distance among the simulated alignment at different times compared to the reference value of the natural alignment, (B) Distribution of pairwise distances for the natural alignment and the last simulation step, (C) Evolution of correlation (spearman/pearson) for one point frequencies between the natural dataset and the simulated one, (D) Artificial (last step) and natural gap distribution over sites.

a background A_i.
The dynamics of pairwise distance in Fig. 3.5 (A) shows that sequences have a lag time to escape the original PSE-1 background.

In this spirit, we can try to classify each site according to its local (CDE) and global (CIE) variability. For the context-dependent entropy we use our model with a PSE-1 background, whereas for context-independent entropy we just compute site entropies of the natural alignment. From Fig. 3.7 (A-B) we see that we can identify three relevant subcategories

- Variable sites (high CIE & CDE) that quickly mutate as they are not constrained by the background
- Conserved sites (low CIE & CDE) that hardly mutate as they are conserved also in the total alignment

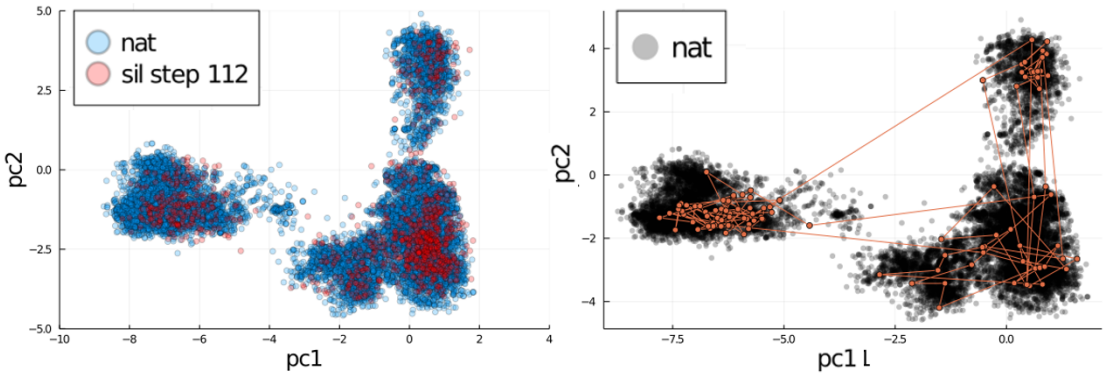

Figure 3.6: Principal component analysis of natural and simulated sequences: (A) Natural sequences compared to the last step of the simulated sequences in their two principal components, (B) comparison of natural sequences (grey) and the trajectoryies of one Markov chain (orange) along the simulation.

- Epistatic sites (high CIE & CDE) which are locally conserved when the context is almost fixed, but as soon as the background moves away from PSE-1 they escape trying to reach their high global variability.

In Fig. 3.7 (C-D) we show the entropy of all sites with a color gradient referring respectively to CDE and CIE. The main trend is that only globally-variable sites rapidly mutate, whereas only highly-conserved positions remain fixed. In addition, in Fig. 3.7 (E) the color gradient relates to |CIE-CDE|, showing that only epistatic sites have a different behaviour with a local and a global regime.
These analysis proves the existence of different timescales related to the cluster exploration. As a matter of fact, the longer timescale that is observed both in variable and conserved sites as a final jump to equilibrium values temporally coincides to the jump of sequences in the PCA space from the cluster of PSE-1 to the other two conglomerations.

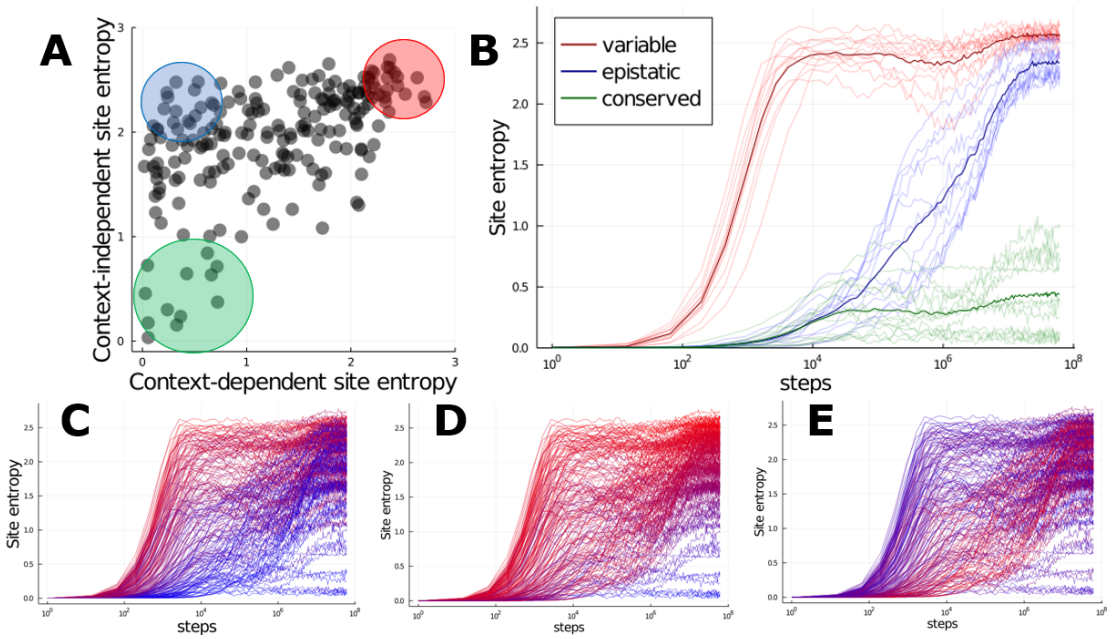

Figure 3.7: Epistatic signals at intermediate scales: (A) Classification of sites in variable, conserved and epistatic according to their CIE (context-independent entropy) and CDE (context-dependent entropy), (B) Evolution of entropies for 12 sites of each category as a function of simulation steps, (C) Evolution of entropies for all sites (color gradient for each site is red/blue for high/low CDE), (D) Evolution of entropies for all sites (color gradient for each site is red/blue for high/low CIE), (E) Evolution of entropies for all sites (color gradient for each site is red/blue for high/low |CIE-CDE|).

Chapter 4

Conclusion

Protein evolution is a complex phenomenon that is driven by several constraints and still needs to be quantitatively explored. This report tried to investigate the topic under the lens of statistical physics inspired data-driven modelling.
The use of computational simulations in the form of a Markov chain occurring in a specific sequence landscape, built from natural homologous protein variants, served as an efficient proxy for describing the mutation-selection dynamics.
The effects of mutations are translated into changes in energy and the selective pressure was modeled through an artificial temperature, with the possibility of reproducing different experimental and environmental conditions.
The sequence alignments produced by sampling the inferred model have effectively captured different stages of natural and in-vitro evolution, encompassing experimental evolution at varying rounds and natural evolution observed in homologous sequences. W are able to connect timescales from few mutations ($\simeq$ 1 year, i.e. the spread of SARS-CoV2) up to distant homology ($\simeq$ 1 billion year, i.e. the timescale of life on earth) in a quantitative way based on distinct kinds of biological data. It is important to underline that the correlation with experimental results is far from obvious and the model had an unexpected high degree of correlation considering that the landscape was learned only on natural sequences that were sufficiently distant from the proteins used in the experiments.
Furthermore, the thesis has explored the emergence of epistatic signals at intermediate scales, providing valuable insights into the important timescales governing the evolutionary dynamics.
A model capturing such fine-scale features in completely different contexts is quite rare: actually all phylogeny and ancestral sequence reconstruction algorithms are based on independent site evolution and therefore neglect epistasis. Our findings contribute to a deeper understanding of the interplay between mutations and selection in shaping protein diversity and the emergence of new protein functions.
Further improvements might consider a more systematic investigation of important

phenomena related to epistasis such as contingency and entrenchment. As a matter of fact, the huge variety of temporal data coming from the simulation dynamics could provide relevant insights about how permissive/restrictive mutations might open/close paths in the evolutionary history.
In addition, the study has highlighted that sites have different dynamics according to their local/global variability. Improvements should look at the dependence of such phenomena on the background initialization.
Moreover, clusters in principal component analysis are related to different protein function: investigating transition paths among subfamilies could help to shed light on emergence of new protein function.
Overall, the development of a quantitative theory of protein evolution could help us to make significant progress in several important tasks such as ancestral sequence reconstruction or protein design and could certainly assist in navigating the immense protein sequence space with a more informed perspective.

Appendix A

Technical details of Direct Coupling Analysis

A.1 Learning Procedure

As we previously highlighted, we aim to learn the probability distribution of the sequence, which translates into quantitatively infer the "fields" $h_i(a_i)$ and the "couplings" $J_{ij}(a_i, a_j)$. Following Maximum Entropy Principle (MEP) [19], we choose a set of observables $O_\alpha(a_1, ..., a_N)$ of the data-set that we want to reproduce with our probabilistic model, that is to say we want:

$$< O_\alpha(a_1, ..., a_N) >_P = \sum_{a_1,...,a_N} P(a_1, ..., a_N) O_\alpha(a_1, ..., a_N) =< O_\alpha >_{Data} \quad \text{(A.1)}$$

These observables will act as constraints on our functional entropy maximization, leading to an expression for the p.d.f. that resembles the common Boltzmann distribution of equilibrium statistical physics.
In our specific case we decide only to impose:

- one-point frequencies $f_i(a) =< \delta_{a,a_i} >_P$
- two-point frequencies $f_{ij}(a, b) =< \delta_{a,a_i}\delta_{b,a_j} >_P$

for every possible i, j, a, b. The justification of such assumption is that we can find several examples in the literature [26, 24, 27] which show that one-site and double-site statistics is sufficient at reproducing higher-order quantities and that the models learned on these two features manage to generate functional sequences. As a matter of fact, it is tempting to add another constraint for example on third-order $f_{ijk}(a_i, a_j, a_k)$ but this is quite risky. If we do some counting, these are 125 terms. The length of the database must be much higher in order to have sufficient statistics, hence introducing such constraint will only make our model learn the noise contained in the data and loose generative power.

A.2 Likelihood maximization

Now that we fixed our starting point, we can show how the algorithm works. Maximizing the entropy of the probability distribution is equal to finding the maximum of the Likelihood in a Bayesian context. As a matter of fact we can exploit Bayes' theorem and write the probability of parameters conditioned to the data as

$$P\left(\{h, \mathbf{J}\} \mid \left\{\mathbf{a^j}\right\}\right) = \frac{P\left(\left\{\mathbf{a^j}\right\} \mid \{h, \mathbf{J}\}\right) P_0(\{h, \mathbf{J}\})}{P\left(\{\mathbf{a^j}\}\right)}, \tag{A.2}$$

where $P_0(\{h, \mathbf{J}\})$ is understood as the prior distribution of the parameters. We can easily implement a gradient ascent on the log-likelihood were the parameters are updated as

$$\begin{aligned} h_i(a) &\leftarrow h_i(a) + \epsilon(f_i(a) - < \delta_{a,a_i} >_P) \\ J_{ij}(a,b) &\leftarrow J_{ij}(a,b) + \epsilon(f_{ij}(a,b) - < \delta_{a,a_i}\delta_{b,a_j} >_P). \end{aligned} \tag{A.3}$$

Obviously, the exact computation of $< \delta_{a,a_i} >_P$ and $< \delta_{a,a_i}\delta_{b,a_j} >_P$ is unfeasible due to the calculation of the partition function. A way out is the approximation of such quantities via Markov Chain Montecarlo sampling (MCMC) which appears to be slow, but leading to correct results. This standard procedure is usually addressed as Boltzmann Machine Learning (bmDCA) [28] and it is currently one of the most widely used techniques in the context of DCA.
If we look back at Eq. (2.2) we realize that the parameters to be inferred scale as $N_{par} = N * q + q^2 \frac{N(N-1)}{2} \simeq O((qN)^2)$ where $q = 21$ is the number of possible amino-acids symbols (20+"-"). Due to this high computational cost, the search for different alternatives or faster implementations is almost necessary. We will investigate possible solutions in the following paragraphs.

A.3 Regularization

The problem that scientists face when studying protein families is that the sequences within these families often do not adhere to the expected patterns. This is due to biased data collection, where the aim is not necessarily to fully explore the sequence space. For example, if researchers are studying a specific protein molecule in birds, they may only sequence it for several different bird species, leading to a family of protein sequences that are similar to each other. This could give the illusion that the probability distribution $P(a_1, ..., a_N)$ of the sequences is concentrated in that region, when in reality it is not.
Additionally, the sequences in a protein family are not independent because they

are linked through their evolutionary history, where mutations have accumulated over time. This is known as sampling bias and is a crucial issue in bio-informatics. In this thesis, we have taken a straightforward approach to address the issue of sampling bias. For the computation of the $f_i(a_i)$ and $f_{ij}(a_i, a_j)$, we have assigned weights to the sequences to reduce the impact of overly similar sequences, effectively addressing the problem of biased data collection:

$$\begin{aligned} f_i(a) &= \frac{\sum_k w_k \delta_{a,a_i^k}}{W_{eff}} \\ f_{ij}(a,b) &= \frac{\sum_k w_k \delta_{a,a_i^k} \delta_{b,a_j^k}}{W_{eff}} \end{aligned} \tag{A.4}$$

where $W_{eff} = \sum_k w_k$ and $k = 1, ..., M$ runs over the sequences of the MSA. How are the weights defined? Since e want to give a smaller importance to sequences that contain many similar copies in the alignment, we decide to put $w_k = \frac{1}{s_k}$ where s_k are the squences that are close to the k-th one. But what does it mean that two sequences are similar? To answer quantitatively, we precisely define the Hamming distance between two sequences **a** and **b**:

$$D_h(\mathbf{a}, \mathbf{b}) = \sum_i (1 - \delta_{a_i,b_i}) = N - \sum_i \delta_{a_i,b_i} \tag{A.5}$$

and we say that s_k is the number of sequences **b** for which $D_h(\mathbf{a}, \mathbf{b}) \geq 0.8 * N$. This means that we consider two proteins as "close" if they share at least 80% of their amino-acids.
In addition, we must deal with another problem that we addressed in the previous section. Since some position in the chain are highly amino-dependent, $f_i(a)$ will be near zero for the majority of amino-acids a. We can notice that almost null frequencies might result in enormously negative local fields. To avoid this result, which is highly non-physical, a pseudo-count is introduced [3]. This consists in adding some constants to empirical one and double point frequencies:

$$\begin{aligned} f_i(a) &\leftarrow (1-\alpha) f_i(a) + \frac{\alpha}{q} \\ f_{ij}(a,b) &\leftarrow (1-\alpha) f_{ij}(a,b) + \frac{\alpha}{q^2}. \end{aligned} \tag{A.6}$$

This procedure is equivalent to adding to the MSA $\alpha/(1-\alpha)$ sequences with residues sampled uniformly. Intuitively, the parameter α should go to zero as the size of the alignment grows to infinity.
Another form of regularization is the introduction of L1 or L2 penalties [3]. This is a common practice in the context of machine learning; the former term forces small

fields and couplings to zero, whereas the latter penalizes large absolute values for these quantities. The net effect is that much sparser networks have to be inferred with a huge gain in speed.

A.4 Training dataset

We take the parameters from the work by Bisardi et al.[25], which were learned on two protein families, PF13354 (Beta-lactamase2) and PF00583 (Acetyltransf1). To generate the MSAs the authors used the hmmsearch command from the HMMer software suite to search the UniProt database. The resulting sequences where then filtered to remove insertions, proteins with more than 10% gaps, and those that were duplicates or closer than 80% to wildtype TEM-1, PSE-1 to avoid the introduction of biases during the bmDCA learning process. Ultimately, the resulting MSAs contained a total of 18,333 homologous and non-identical aligned sequences of length 202 for PF13354 and 43,576 for PF00583. The parameter inference was then carried on by using a Potts model with bmDCA [29], that exploits the most precise DCA models. Our analysis will only limit to the use of the parameters related to PF13354 family.

Appendix B

Note on MCMC sampling

B.1 Equilibrium distribution

Consider an amino-acid sequence $A = (a_1, \cdots, a_L)$, with $a_i \in \{1, \cdots, q = 21\}$ including gaps, its energy $H(A)$ and a Boltzmann model

$$P(A) = \frac{e^{-\beta H(A)}}{Z} , \qquad Z = \sum_A e^{-\beta H(A)} . \tag{B.1}$$

Now consider a nucleotide sequence $B = (b_1, \cdots, b_{3L})$ with $b_i \in \{1,2,3,4\}$, such that

- each codon codes for an amino acid, e.g. $(b_1, b_2, b_3) \to a_1$, etc., with the exception of the stop codons, which have then to be excluded (see below);
- we call $A(B)$ the amino acid sequence corresponding to B;
- we note that $A(B)$ has no gaps because no codon codes for a gap; if we want to code for gaps we could add a 5th symbol $b_i = 5$ to the nucleotide alphabet and assume for example than any codon containing a 5 codes for a gap; however, we will exclude gaps (see below).

We can now associate the energy $H[A(B)]$ to each sequence, and we call $\mathcal{B}(A) = |\{B : A(B) = A\}|$ the number of sequences B that correspond to the same A. Because each codon independently codes for an amino acid, we have

$$\mathcal{B}(A) = \prod_{i=1}^{L} \mathcal{B}(a_i) , \tag{B.2}$$

where $\mathcal{B}(a_i)$ is the number of codons that code for a_i. We assign to a nucleotide sequence the probability

$$P(B) = \frac{1}{\mathcal{B}(A(B))} \frac{e^{-\beta H[A(B)]}}{Z} = \frac{e^{-\beta \mathcal{H}(B)}}{Z} , \tag{B.3}$$

having introduced a modified Hamiltonian

$$\mathcal{H}(B) = H[A(B)] + T \log \mathcal{B}(A(B)) = H[A(B)] + T \sum_{i=1}^{L} \log \mathcal{B}[a_i(B)] \ . \tag{B.4}$$

In this way, the probability of a sequence A and the partition function are the same as that of the original model:

$$Z = \sum_B e^{-\beta \mathcal{H}(B)} = \sum_A e^{-\beta H(A)} \sum_{B:A(B)=A} \frac{1}{\mathcal{B}(A(B))} = \sum_A e^{-\beta H(A)} \ . \tag{B.5}$$

So, we have a reweighting of nucleotide sequences, with lower weight/higher energy being assigned to sequences that code for proteins that have more possible representations in the genetic code. This takes into account the fact that different codons can code the same aminoacids and favors with a higher probability/lower energy the sequences that have a low degeneracy, i.e. that can be coded by just a few combinations of nucleotides. Instead, amino-acids that have a high genetic redundancy are more rarely sampled. In this way, probability is correctly distributed among all possible codons. One assumption is that we do not have codon biases: the energy is defined only on amino-acid, it is blind to synonymous codons that are translated into the same amino-acid.

B.2 Metropolis sampling

Now, we want to set up a Metropolis sampling on the codon space. We introduce symmetric transition probabilities for a single codon $\Omega(b \to b') = \Omega(b' \to b)$ with $\sum_{b'=1}^{4} \Omega(b \to b') = 1$. We call B the sequence with b and B' the sequence with $b \to b'$. The Metropolis acceptance rate is

$$p(b \to b') = \min\left(1, e^{-\beta[\mathcal{H}(B') - \mathcal{H}(B)]}\right) = \min\left(1, \frac{\mathcal{B}(A(B))}{\mathcal{B}(A(B'))} e^{-\beta[H[A(B')] - H[A(B)]]}\right) \ . \tag{B.6}$$

The Metropolis algorithm proceed as follows:

- Choose a codon $b = (b_1, b_2, b_3)$ at random.
- Propose a change to a new codon b' reachable from b through just one aminoacid substitution with probability $\Omega(b \to b')$.
- Accept the change with probability $p(b \to b')$.

This process satisfies detailed balance with respect to $p(B)$ because

$$p(B)\Omega(b_i \to b_i')p(b_i \to b_i') = p(B')\Omega(b_i' \to b_i)p(b_i' \to b_i) \ . \tag{B.7}$$

Furthermore, and most importantly, we can restrict the space of possible moves by setting some of the $\Omega(b_i \to b'_i)$ to zero. The only crucial requirement is that $\Omega(b \to b')$ is symmetric. In our case we chose the following proposal matrix to model the insertion/deletion process:

$$\Omega(b \to b') = \begin{cases} 0, & \text{if } b, b' = G \\ \frac{1}{64}, & \text{if } b = G, b' = C_i \text{or} b = C_i, b' = G \\ \frac{63}{64}\delta_{ij}, & \text{if } b = C_i, b' = C_j \end{cases}$$

where G is a gap and C_i is one among the possible 64-codons observed in nature. If some moves are forbidden, the resulting MC chain is not ergodic, and it will remain confined in the space of allowed sequences, which we call $\boldsymbol{B}$. Still the probability of a sequence in equilibrium is

$$P(B) = \frac{e^{-\beta \mathcal{H}(B)}}{Z'} \ , \tag{B.8}$$

with a modified partition function

$$Z' = \sum_{B \in \boldsymbol{B}} e^{-\beta \mathcal{H}(B)} \ . \tag{B.9}$$

Let us assume that the forbidden states are chosen carefully in such a way that for each allowed amino acid sequence $A \in \boldsymbol{A}$ we have the same number $\mathcal{B}(A)$ of coding nucleotide sequences. Then we still have

$$Z' = \sum_{B \in \boldsymbol{B}} e^{-\beta \mathcal{H}(B)} = \sum_{A \in \boldsymbol{A}} e^{-\beta H(A)} \sum_{B \in \boldsymbol{B} : A(B) = A} \frac{1}{\mathcal{B}(A(B))} = \sum_{A \in \boldsymbol{A}} e^{-\beta H(A)} \ . \tag{B.10}$$

B.3 Gibbs Sampling

In addition, we can also set up a Gibbs sampling on the nucleotide space.
We define the probability of a single nucleotide mutation, conditional to all other nucleotides as $P(b_{i_k}|B_{-i_k})$, where

- b_{i_k} is the nucleotide in codon i taking position $k \in \{1,2,3\}$
- B_{-i_k} stands for the sequence with the removal of that specific nucleotide
- $\boldsymbol{b} = \{A, C, G, T\}$ is the single nucleotide space

This probability can be expressed as:

$$P(b_{i_k}|B_{-i_k}) = \frac{P(B)}{P(B_{-i_k})} \tag{B.11}$$

where $P(B)$ is the usual as in Eq. (B.3) and $P(B_{-i_k}) = \sum_{b_{i_k} \in \boldsymbol{b}} P(B)$. If we compute it explicitly we obtain:

$$P(b_{i_k}|B_{-i_k}) = \frac{\frac{e^{-\beta H[A(B)]}}{\mathcal{B}(A(B))}}{\sum_{b_{i_k} \in \boldsymbol{b}} \frac{e^{-\beta H[A(B)]}}{\mathcal{B}(A(B))}} \tag{B.12}$$

and we see that the advantage of this procedure is that we don't need anymore the partition function.
The Gibbs algorithm works iteratively in the following way:

- Choose a nucleotide i_k at random among the non-gapped positions
- Compute the probability of all possible mutations in that position $P(b'_{i_k}|B_{-i_k})$
- Sample the new nucleotide b'_i according to the probability density $P(b'_{i_k}|B_{-i_k})$

If we assume our hamiltonian to take the form of a Potts hamiltonian, that is to say:

$$H = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j) \tag{B.13}$$

with a_i the aminoacid coded by the codon i containing b_{i_k}, then $P(b_{i_k}|B_{-i_k})$ reads as follows:

$$P(b_{i_k}|B_{-i_k}) = \frac{e^{\beta h_i(a_i) + \beta \sum_j J_{ij}(a_i,a_j)} (\mathcal{B}(a_i))^{-1}}{\sum_{b_{l_k} \in \boldsymbol{b}} e^{\beta h_l(a_l) + \beta \sum_j J_{lj}(a_l,a_j)} (\mathcal{B}(a_l))^{-1}} \tag{B.14}$$

with the caveat that the nucleotide dependence is hidden in the aminoacid term $a_l = a_l(b_{l_1}, b_{l_2}, b_{l_3})$. We notice that such probability density is much faster to obtain, since we reduced the initial intensive computation to just a few terms to evaluate.

Acknowledgements

Firstly, I would like to thank my supervisors Martin and Francesco. Their expertise and help has been crucial in the realization of this work. I have relevantly broadened my knowledge in these months thanks to their suggestions and constant availability for discussions. I must thank them also for having always enhanced a stimulating environment in the lab.
Secondly, I surely have to thank Matteo for his longstanding presence during my internship. He guided me through the different tasks always showing empathy and support. I would like to extend my gratitude also to the rest of the group which positively created a familiar learning environment.
Now, I cannot forget to mention my classmates, especially the ones living with me in Trieste. If I have completed this master it is also thanks to them who supported me during the worst moments.
I must thank all my friends scattered around France and Italy living in Perugia, Trieste, Turin and Paris. At each step of my journey they have been the sign of a positivity that will always try to reach me wherever I will be.
Furthermore, I would like to express my heartfelt thanks to my family that has constantly stood by me non-regarding of the distance and the difficulties.
Lastly, my final thanks goes to Claudia. I owe so many things to her love and support. Her presence has been a source of unwavering joy and these two years have only confirmed how grateful I am of having encountered her.

Acronyms

AI

artificial intelligence

DCA

Direct Coupling Analysis

MEP

Maximum Entropy Principle

bmDCA

Boltzmann Machine Direct Coupling Analysis

plmDCA

Pseudo-likelihood Maximization Direct Coupling Analysis

Bibliography

[1] William P. Russ et al. «An evolution-based model for designing chorismate mutase enzymes». In: *Science* 369.6502 (2020), pp. 440–445. DOI: 10.1126/science.aba3304. eprint: https://www.science.org/doi/pdf/10.1126/science.aba3304. URL: https://www.science.org/doi/abs/10.1126/science.aba3304 (cit. on p. 1).

[2] Faruck Morcos et al. «Direct-coupling analysis of residue coevolution captures native contacts across many protein families». In: *Proceedings of the National Academy of Sciences* 108.49 (2011), E1293–E1301. DOI: 10.1073/pnas.1111471108. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1111471108. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1111471108 (cit. on p. 1).

[3] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. «Inverse statistical physics of protein sequences: a key issues review». In: *Reports on Progress in Physics* 81.3 (Jan. 2018), p. 032601. DOI: 10.1088/1361-6633/aa9965. URL: https://dx.doi.org/10.1088/1361-6633/aa9965 (cit. on pp. 1, 8, 10, 29).

[4] Michael A. Stiffler et al. «Protein Structure from Experimental Evolution». In: *Cell Systems* 10.1 (2020), 15–24.e5. ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2019.11.008. URL: https://www.sciencedirect.com/science/article/pii/S2405471219304284 (cit. on pp. 1, 5–7, 15, 20).

[5] Marco Fantini, Simonetta Lisi, Paolo De Los Rios, Antonino Cattaneo, and Annalisa Pastore. «Protein Structural Information and Evolutionary Landscape by In Vitro Evolution». In: *Molecular Biology and Evolution* 37.4 (Oct. 2019), pp. 1179–1192. ISSN: 0737-4038. DOI: 10.1093/molbev/msz256. eprint: https://academic.oup.com/mbe/article-pdf/37/4/1179/32960043/msz256.pdf. URL: https://doi.org/10.1093/molbev/msz256 (cit. on pp. 1, 5, 7, 15).

[6] Christian B. Anfinsen. «Principles that Govern the Folding of Protein Chains». In: *Science* 181.4096 (1973), pp. 223–230. DOI: 10.1126/science.181.4096.223. eprint: https://www.science.org/doi/pdf/10.1126/science.181.

4096.223. URL: https://www.science.org/doi/abs/10.1126/science.181.4096.223 (cit. on p. 2).

[7] David Penny. «Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts.» In: *Systematic Biology* 53.4 (Aug. 2004), pp. 669–670. ISSN: 1063-5157. DOI: 10.1080/10635150490468530. eprint: https://academic.oup.com/sysbio/article-pdf/53/4/669/24197744/53-4-669.pdf. URL: https://doi.org/10.1080/10635150490468530 (cit. on p. 3).

[8] Yeonwoo Park, Brian P. H. Metzger, and Joseph W. Thornton. «Epistatic drift causes gradual decay of predictability in protein evolution». In: *Science* 376.6595 (2022), pp. 823–830. DOI: 10.1126/science.abn6895. eprint: https://www.science.org/doi/pdf/10.1126/science.abn6895. URL: https://www.science.org/doi/abs/10.1126/science.abn6895 (cit. on p. 5).

[9] David McCandlish, Etienne Rajon, Premal Shah, Yang Ding, and Joshua Plotkin. «The role of epistasis in protein evolution». In: *Nature* 497 (May 2013), E1–E2. DOI: 10.1038/nature12219 (cit. on p. 5).

[10] Tyler N Starr and Joseph W Thornton. «Epistasis in protein evolution». In: *Protein science* 25.7 (2016), pp. 1204–1218 (cit. on p. 5).

[11] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. «Correlated mutations and residue contacts in proteins». In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994), pp. 309–317. DOI: https://doi.org/10.1002/prot.340180402. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340180402. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340180402 (cit. on p. 8).

[12] E Neher. «How frequent are correlated changes in families of protein sequences?» In: *Proceedings of the National Academy of Sciences* 91.1 (1994), pp. 98–102. DOI: 10.1073/pnas.91.1.98. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.91.1.98. URL: https://www.pnas.org/doi/abs/10.1073/pnas.91.1.98 (cit. on p. 8).

[13] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511790492 (cit. on p. 9).

[14] Robert D. Finn et al. «The Pfam protein families database: towards a more sustainable future». In: *Nucleic Acids Research* 44.D1 (Dec. 2015), pp. D279–D285. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1344. eprint: https://academic.oup.com/nar/article-pdf/44/D1/D279/9484040/gkv1344.pdf. URL: https://doi.org/10.1093/nar/gkv1344 (cit. on p. 9).

[15] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. «Correlated mutations and residue contacts in proteins». In: *Proteins: Structure, Function, and Bioinformatics* 18.4 (1994), pp. 309–317. DOI: https://doi.org/10.1002/prot.340180402. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340180402. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340180402 (cit. on p. 9).

[16] E Neher. «How frequent are correlated changes in families of protein sequences?» In: *Proceedings of the National Academy of Sciences* 91.1 (1994), pp. 98–102. DOI: 10.1073/pnas.91.1.98. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.91.1.98. URL: https://www.pnas.org/doi/abs/10.1073/pnas.91.1.98 (cit. on p. 9).

[17] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. «Inverse statistical problems: from the inverse Ising problem to data science». In: *Advances in Physics* 66.3 (2017), pp. 197–261. DOI: 10.1080/00018732.2017.1341604. eprint: https://doi.org/10.1080/00018732.2017.1341604. URL: https://doi.org/10.1080/00018732.2017.1341604 (cit. on p. 10).

[18] Yasser Roudi, Joanna Tyrcha, and John Hertz. «Ising model for neural data: Model quality and approximate methods for extracting functional connectivity». In: *Phys. Rev. E* 79 (5 May 2009), p. 051915. DOI: 10.1103/PhysRevE.79.051915. URL: https://link.aps.org/doi/10.1103/PhysRevE.79.051915 (cit. on p. 10).

[19] E. T. Jaynes. «Information Theory and Statistical Mechanics». In: *Phys. Rev.* 106 (4 May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620 (cit. on pp. 10, 11, 27).

[20] E I Shakhnovich and A M Gutin. «Engineering of stable and fast-folding sequences of model proteins.» In: *Proceedings of the National Academy of Sciences* 90.15 (1993), pp. 7195–7199. DOI: 10.1073/pnas.90.15.7195. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.90.15.7195. URL: https://www.pnas.org/doi/abs/10.1073/pnas.90.15.7195 (cit. on p. 11).

[21] Eugene I Shakhnovich. «Protein design: a perspective from simple tractable models». In: *Folding and Design* 3.3 (1998), R45–R58. ISSN: 1359-0278. DOI: https://doi.org/10.1016/S1359-0278(98)00021-2. URL: https://www.sciencedirect.com/science/article/pii/S1359027898000212 (cit. on p. 11).

[22] William P Russ, Drew M Lowery, Prashant Mishra, Michael B Yaffe, and Rama Ranganathan. «Natural-like function in artificial WW domains». In: *Nature* 437.7058 (Sept. 2005), pp. 579–583. ISSN: 0028-0836. DOI: 10.1038/nature03990. URL: https://doi.org/10.1038/nature03990 (cit. on p. 11).

[23] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. «Evolutionary information for specifying a protein fold». In: *Nature* 437.7058 (Sept. 2005), pp. 512–518. ISSN: 0028-0836. DOI: 10.1038/nature03991. URL: https://doi.org/10.1038/nature03991 (cit. on p. 11).

[24] William P. Russ et al. «An evolution-based model for designing chorismate mutase enzymes». In: *Science* 369.6502 (2020), pp. 440–445. DOI: 10.1126/science.aba3304. eprint: https://www.science.org/doi/pdf/10.1126/science.aba3304. URL: https://www.science.org/doi/abs/10.1126/science.aba3304 (cit. on pp. 11, 27).

[25] Matteo Bisardi, Juan Rodriguez-Rivas, Francesco Zamponi, and Martin Weigt. «Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution». In: *Molecular Biology and Evolution* 39.1 (Nov. 2021). msab321. ISSN: 1537-1719. DOI: 10.1093/molbev/msab321. eprint: https://academic.oup.com/mbe/article-pdf/39/1/msab321/46861804/msab321.pdf. URL: https://doi.org/10.1093/molbev/msab321 (cit. on pp. 15, 17, 19, 30).

[26] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. «Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis». In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12186–12191. DOI: 10.1073/pnas.1607570113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1607570113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1607570113 (cit. on p. 27).

[27] Kai Shimagaki and Martin Weigt. «Selection of sequence motifs and generative Hopfield-Potts models for protein families». In: *bioRxiv* (2019). DOI: 10.1101/652784. eprint: https://www.biorxiv.org/content/early/2019/09/05/652784.full.pdf. URL: https://www.biorxiv.org/content/early/2019/09/05/652784 (cit. on p. 27).

[28] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. «A learning algorithm for boltzmann machines». In: *Cognitive Science* 9.1 (1985), pp. 147–169. ISSN: 0364-0213. DOI: https://doi.org/10.1016/S0364-0213(85)80012-4. URL: https://www.sciencedirect.com/science/article/pii/S0364021385800124 (cit. on p. 28).

[29] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. «How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?» In: *Molecular Biology and Evolution* 35.4 (Jan. 2018), pp. 1018–1027. ISSN: 0737-4038. DOI: 10.1093/molbev/msy007. eprint: https://academic.oup.com/mbe/article-pdf/35/4/1018/24597926/msy007.pdf. URL: https://doi.org/10.1093/molbev/msy007 (cit. on p. 30).