POLYTECHNIC UNIVERSITY OF TURIN

Master's Degree in Physics of Complex Systems

Master's Degree Thesis

Towards statistical-physics inspired modeling of experimental protein evolution

Supervisors

prof. Alfredo BRAUNSTEIN

prof. Martin WEIGT

prof. Francesco ZAMPONI

Candidate

Matteo BISARDI

July 2020

Contents

1 Introduction and motivation

Data-driven modelling approaches, including those inspired by statistical physics of complex and disordered systems, are rapidly gaining importance in modern computational biology. In this thesis we propose approaches to the modelling of experimental evolution protocols like directed evolution.The latter proceed by alternating cycles of mutation (e.g. by error-prone polymerase chain reaction) and selection for any function (e.g. for antibiotic resistance or other enzymatic activity) for some protein of interest.

Recently it has been shown in two independent articles [1, 2] that sequence ensembles generated by this approach can be used to gain important structural and functional information about the studied proteins. However, the basic understanding of the potential and the limitations of the experimental approaches remains limited, and the reasons leading to significant differences between the two sets of experimental results remain unclear.

Here we address this question from a statistical-physics inspired point of view. We explore data-driven sequences landscapes inferred using a method from inverse statistical physics called Direct-Coupling Analysis (DCA), which infers Potts models from multiple-sequence alignments of natural proteins via a maximum-entropy approach.

We first show that sequence data coming from the two evolution experiments could be well described by the DCA sequence landscape, and that the experimental sequences are coherent with sampling of sequence space using Gibbs sampling. The results is highly non trivial in that DCA sequence landscapes, which are learned on naturally evolved sequences, are selected for a number of naturally occurring phenotypes, and yet they are able to describe in vitro evolution, in which a specific human-designed phenotype is used for selection.

Exploiting these results, we can simulate in silico evolution protocols, which allow us to assess systematically important experimentally-tuned characteristics like the number of mutations per sequence, the strength of selection and the number of analysed sequences and consequently unveil their role for protein structure prediction. We show that this analysis can explain the different performances of the two recently published experimental works. It also opens the possibility to provide a priori estimates of the optimal value of experimental parameters, in order to optimise the experimental protocols.

We also anticipate that a further refinement of our modelling approaches can be obtained by a more realistic description of the evolutionary dynamics, e.g. by modelling mutations at the level of the genetic DNA sequence instead of the protein's amino-acid sequence. This would allow to take into account details of the experiments (e.g. DNA mutational biases) and open the way to an intense exchange between models and experiments.

2 Direct Coupling Analysis and applications

In the following we introduce the Direct Coupling Analysis (DCA) [3] approach. The aim of DCA is to construct a probabilistic model to describe protein sequences exploiting statistical features extracted from large biological databases. DCA assigns a probability score $P(\underline{A})$ to every sequence $\underline{A}$ of amino-acids in a protein family, from which relevant biological information can be retrieved: examples include residue-residue contact prediction [4], prediction of fitness effect of mutations [5] and artificial protein design [6]. However, before entering into the description of DCA, we will recall a few facts and notations about proteins and their evolution, which justify the application of inverse statistical physics to protein sequence ensembles.

2.1 A little bit of biology

Proteins

Proteins are biological molecules necessary for the functioning of almost all cellular processes. They belong to the most fascinating complex systems in nature, enormous progress has been made over the years to understand them, with incredible scientific insights and fantastic applications. But still we miss very basic elements, and important limitations (e.g. in terms of protein dynamics or sequence-function mapping) persist.

Each protein is primarily a one dimensional polymer composed by amino acids that folds into a convoluted shape, the folded protein, by bringing amino acids separated by a long distance along the linear sequence into close physical proximity.

The three-dimensional structure of proteins in most cases is tightly related to their function, the knowledge of the first is key to understand - and potentially engineer - the latter. It was speculated long ago that the structure of proteins depends only upon their amino acid sequence [7], that is the chain of monomers that constitutes them. As a consequence the full 3D structure of proteins is encoded in a single sentence: its 20-letter amino acid sequence, a pictorial representation is shown in Fig.1. Over the last decades, big efforts had been made in the field of biological physics to decode this relationship, that is to solve the protein-folding problem.

Innovations and discoveries in this field would allow for novel applications in medicine and biotechnology. For example, the ability to design new protein folds [8] would make possible the exploration of large regions of the protein universe not yet observed in nature, allowing for precise engineering at the molecular scale.

Recently, a new tool, Direct Coupling Analysis [9], has emerged in the field of protein structure prediction exploiting the fact that proteins conserve their 3D structure and biological function throughout evolution, while substituting up to 70-80% of their amino acids.

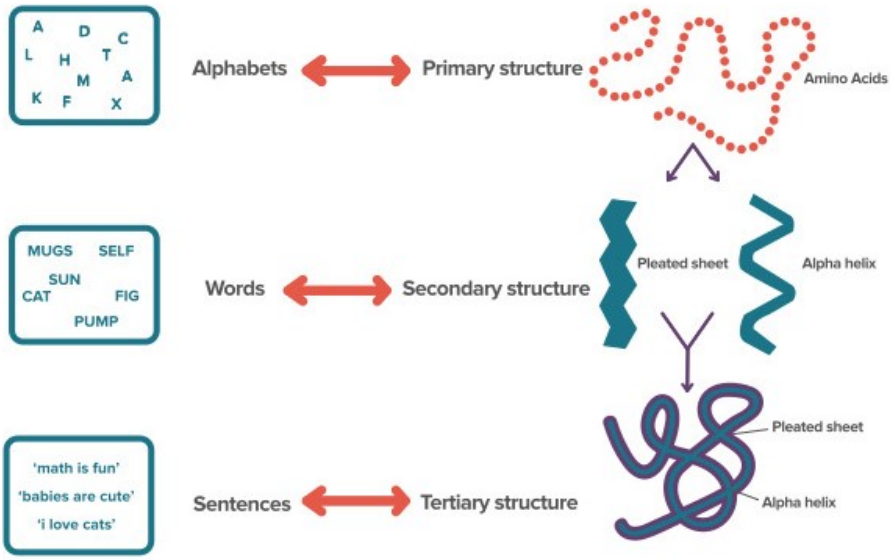


Figure 1: Intuitive representation of the hierarchical level of structural organisation in proteins compared to that of language. Figure source: internet, unknown author

Protein families

In the course of organisms evolution, proteins evolve as well. Mutations at DNA level in the form of nucleotide changes or insertions and deletions translate into mutations at amino acid level. Even if function and biological activity of a protein change little during evolution, different organisms can accumulate mutations and now present a high variety in amino acid sequences.

As a consequence we now observe across all domains of life many different sequences that present the same structure and leave the functionality of the protein fundamentally impaired. It is natural to group proteins which have similar structure and function, into labelled collections and consider them as variants of the same protein. The set of such sequences makes a protein family.

Thanks to the recent revolution in sequencing technology, a lot of sequences have become available and are accumulating at exponential speed. For example, the PFAM [10] database contains protein sequences subdivided by domains that evolve and fold almost independently with respect to the rest of the protein. Many of the more than 18000 PFAM protein families are large and contain more than 1000 sequences, an invaluable source for data-driven modelling approaches like DCA.

2.2 The model

DCA is based on two crucial features: on one side protein sequences can be organized into families that usually share very similar three dimensional structures and biolog-

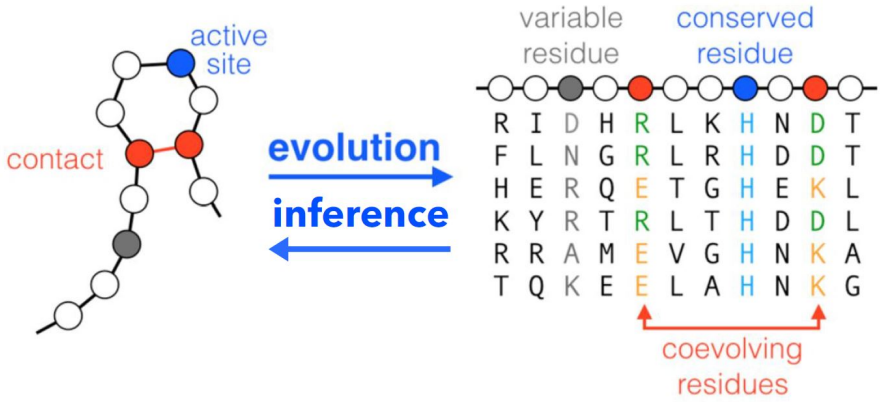


Figure 2: A pictorial representation of the MSA of a family of proteins and its relationship with the co-evolution of amino acids pairs.

ical activity, on the other side those sequences present a very high inter-variability in terms of amino acids.

DCA aims at exploiting this variability in the amino acid sequences to implicitly learn the constraints - often related to the folding into a specific 3D structure - that statistically characterize those proteins (Fig.2).

Proteins belonging to the same family are characterized by patters of conservation in their amino acid sequences. Some residues are amino acid specific, that is only one precise amino acid is ever found at that position in natural sequences, cf. the blue column in Fig.2. This is typically due to local (binding, functional) or global (folding, stability) constraints. Such patterns have long been exploited to build models for protein families with a discrete success.

DCA exploits those pattern of conservation, but integrates also information related to another common feature in sequences belonging to the same family: patterns of correlated amino acid substitution at specific sites, cf the red positiond in Fig.2.

The basic hypothesis connecting amino acid substitution patterns in sequences and residue–residue contacts is simple: in case of a pair of residues being in contact in the folded state, a substitution of an amino acid at one position with destabilizing effects would be compensated by a substitution of the other position over the evolutionary timescale, in order for the pair to maintain a similar physical interaction.

Multiple Sequence Alignments

Sequences belonging to the same family can be gathered in a Multiple Sequence Alignment (MSA), a matrix of $M \times N$ letters whose M rows represent the sequences of amino acids $(a_1, ..., a_N)$ of distinct proteins of N sites. Each of the letters a_n^m, $n = 1, .., N \;\; m = 1, ..., M$, belongs to an alphabet of 21 symbols: the 20 letters that typically represent amino acids and an extra symbol, a gap " $-$ ", used to deal with insertions and deletions in individual proteins. For practical reasons concerning the

algorithmic implementation of DCA those amino acids are usually mapped into the numbers from 1 to 21. We will use the same notation throughout this work. The length of a typical PFAM domain family typically ranges from 50 to 500 sites and the number of sequences in the same family can reach up to $M = 10^6$.

The MSA of a protein family is characterized by the presence of specific statistical patterns in the amino acid letters it contains, namely within and between the columns of the matrix. The simplest statistical features are given by single site and double site empirical amino acid frequencies. The MSA of a specific protein family is characterized by some very conserved sites, that is specific positions in the proteins (or equivalently in the columns of the matrix), that mostly contain the same amino acid (same symbol in the matrix).

Those very conserved sites are useful to build MSAs in the first place, that is to find and align protein sequences that belong to the same family. Moreover those very conserved sites suggest that the amino acid that they present more often (as amino acid H in Fig.2) is crucial to perform protein function, indeed during evolution it has almost never changed despite great variability in other sites.

Another interesting feature, as already mentioned, is given by patterns of 2-site correlations, that is couples of sites $i, j = 1, ..., N$ that present couples of amino acids (a_i, a_j) that are likely to be present together in the same protein. A pictorial example is shown in Fig.2 with amino acids (R, D) and (E, K).

Single and two site statistics seems to carry too little information to fully model a whole protein family. Nevertheless in a seminal work Ranganathan and coworkers [11, 12] showed that the pattern of pairwise residue co-variation was sufficient to generate new functional protein sequences. Even more interesting was the fact that only taking into account single-column statistics (of the MSA matrix) resulted in non-functional amino acid sequences.

Global inference and Potts models

To quantify and understand two-site correlations, measures like mutual information and co-variance have been used with some success to elucidate protein-protein interactions in bacterial two-component signaling pathways [13, 14]. Nonetheless such local correlation methods lack the ability to disentangle direct from indirect correlations. An example is given by the case where residue i is coupled directly to j, and j to k. Then i and k will also show some correlation, without being directly coupled. The effect could be amplified if multiple paths of couplings connecting i and k are present.

A global statistical modelling approach (DCA) was then proposed by Weigt et al. [3, 4] to model MSA of protein families. The hope is that such a global modelling approach would be able to disentangle direct from indirect interactions. The basic assumption of modeling MSAs using methods from (inverse) statistical physics [15] is that protein sequences from the same family represent i.i.d. samples drawn from a Boltzmann distribution:

$$P(a_1, ..., a_N) = \frac{e^{-H(a_1,...,a_N)}}{Z} \tag{1}$$

where the the inverse temperature β commonly used in statistical mechanics is conventionally set to 1. The Hamiltonian reads:

$$H(a_1, ..., a_N) = -\sum_{i}^{N} h_i(a_i) - \sum_{i \leq j}^{N} J_{ij}(a_i, a_j) \tag{2}$$

and the partition function Z

$$Z = \sum_{a_1=1}^{21} ... \sum_{a_N=1}^{21} e^{-H(a_1,...,a_N)} \tag{3}$$

The Hamiltonian H - in statistical mechanics terms - represents the energy of the system. Here it does represent a statistical prevalence, but we use the term "energy" by analogy.

The form of the Hamiltonian (2) is borrowed from statistical mechanics: it is typically used to describe the energy of a system of interacting spins, called Potts model. Potts models are an extension of Ising models, where the number of possible states that a spin can take (referred as q) is more than 2 and in this case corresponds to number of distinct symbols in the MSA, i.e. $q = 21$.

In the following we will refer to h_i as "fields" and to J_{ij} as "couplings", borrowing again the terms from statistical mechanics. Each protein residue is interpreted as a spin of a magnetic systems with categorical values: the $q = 21$ amino acids. In this specific Potts model fields and couplings depend upon the amino acid present at the site, hence for each site i the field is a vector of length q and J_{ij} a $q \times q$ matrix. Within this interpretation a big field $h_i(a_i = a)$ means that amino acid a is very likely to be in position i, hence the site is a conserved one. In the same spirit a big positive couplings $J_{ij}(a_i = a, a_j = b)$ represents an attractive interaction between the two spins, thereby favouring the contextual presence of amino acids a at site i and amino acid b at site j.

The question that now arises naturally is: is it justified to use Potts-like describe protein families? First of all, we can observe that when the couplings are all equal to zero, the model reduces to a profile, or site-independent model. Those models already belong to the most successful tools in bioinformatics; they are at the basis of most techniques for multiple-sequence alignment and homology detection [16]. When non-zero couplings are present, the hope is that those couplings reflect interactions that are biologically interpretable, such as structural proximity of the corresponding residues.

Another argument in favour of this functional form for the probability distribution is given by the Maximum Entropy Principle (MEP), first introduced by Jaynes [17]. MEP is a powerful tool to come up with statistical models in absence of

information about the system to be modelled. It can be shown with simple analytical computations that the least constrained probability distribution that is able to reproduce single and two-site statistics of the MSA is exactly the Boltzmann distribution of eq. 1.

Inverse statistical physics and learning

Given this choice for the model, fields and couplings still need to be determined for each protein family. How? This is a task for inverse statistical physics.

The central goal of statistical physics is to derive the mean observable quantities of a system from the physical interactions of its constituents. In the case of the Ising model, one starts by defining the interaction between its microscopic constituents, the spins, and then derives mean observables, like the magnetisation. In an inverse problem, instead, the starting point are the observations of the configurations of a system whose parameters are unknown and the goal is to figure out those parameters. In the context of our specific problem the unknowns are the fields and couplings and the observations are given by the amino acid sequences present in the protein families.

In particular the quantities that we observe and that we want the model to fit are the frequency of occurrences $f_i(a)$ of amino acid a in column i, and the co-occurency $f_{ij}(b,c)$ of amino acids b and c in columns i and j computed from the MSA of the protein family

$$f_i(a) = \frac{1}{M_{eff}} \sum_{m=1}^{M} w_m \delta_{a,a_i^m} \tag{4}$$

$$f_{i,j}(b,c) = \frac{1}{M_{eff}} \sum_{m=1}^{M} w_m \delta_{b,a_i^m} \delta_{c,a_j^m} \tag{5}$$

where M_{eff} and w_m are used to account for sampling biases. Those biases occur because sequence databases often contain many sequences very similar to each other, given that they come from evolutionary close species. To define a reweighing procedure that allows us to treat all samples of a MSA approximately as independent, we employ a simple metric, the Hamming distance $d_H(\underline{a}, \underline{b})$ between a pair of sequences $\underline{a}, \underline{b}$. Given that all the sequences in the dataset have the same length, this distance is a properly defined metric:

$$d_H(\underline{a}, \underline{b}) = \sum_{i=1}^{N} (1 - \delta_{a_i,b_i}) = N - \sum_{i=1}^{N} \delta_{a_i,b_i} \tag{6}$$

The weight w_m of a sequence is defined as the inverse of number of sequences (itself included) that share more than $\gamma = 0.8 \cdot N$ identity with $\underline{a}$. The effective number

of sequences in a MSA M_{eff} is then defined as

$$M_{eff} = \sum_{m=1}^{M} w_m \tag{7}$$

Given a guess for the parameters h_i, J_{ij}, to verify if our model fits correctly this statistics, we would need to compute marginals of the model, which has the same computational complexity as the calculation of the partition function. The partition function involves a sum with q^N terms, thereby making the task computationally unfeasible. Approximation methods have been developed to deal with this problem, from message passing [3] and mean-field [4] approaches - originally developed to solve direct problems in statistical mechanics (i.e. from fields and couplings to marginals) - to approximate Bayesian inference methods like pseudo-likelihood maximisation [18].

Although being very fast (they can run on a personal computer in matter of minutes), they lack generative power, that is, beyond the quantities that they are required to approximate, they fail to capture higher order statistics.

A more precise, though slower method, to infer the parameters of the Potts model is Boltzmann machine learning [19]. Starting from an initialisation of the parameters, 1-point and 2-point marginals $P(a_i = a)$, $P(a_i = b, a_j = c)$ are estimated via Monte Carlo Markov Chain sampling. The results are then compared with the empirical marginals $f_i(a), f_{ij}(b, c)$ and the parameters are updated according to the following rules:

$$\begin{aligned} h_i(a)^{new} &= h_i(a)^{old} + \epsilon(f_i(a) - P(a_i = a)) \\ J_{ij}^{new} &= J_{ij}^{old} + \epsilon(f_{ij}(b, c) - P(a_i = b, a_j = c)) \end{aligned} \tag{8}$$

This procedure is iterated, until empirical and model distribution are coherent. In this case of very large MCMC samples and many iterations, this method is guaranteed to converge to the exact solution.

2.3 Biological applications

Suppose that we have very carefully estimated the parameters h, J of the Potts Hamiltonian. What information about the proteins in the protein family can we extract from them?

Residue contact prediction

The most successful task of DCA so far has been residue-residue contact prediction [20]. This task is considered hard, as already mentioned, and prior to this technique no computationally efficient and accurate methods were available. The basic idea is quite simple: using the couplings J_{ij} as a proxy for physical interaction of the

residues in the 3D folded structure. A strong couplings between i, j would indicate a strong evolutionary constraint, itself caused by the physical proximity of the residues.

One of the first scores for physical proximity of residues i, j that can be extracted from the $q \times q$ matrix $J_{ij}(a, b)$ is the Frobenius norm [21]:

$$F_{ij} = \sqrt{\sum_{b,c=1}^{q} J_{ij}^2(b, c)} \tag{9}$$

The sites predicted to be in contact will be those with the highest F_{ij} values. An empirically even better score can be achieved implementing an average-product correction (APC) [22]:

$$F_{ij}^{APC} = F_{ij} - \frac{\sum_l F_{il} \sum_k F_{kj}}{\sum_{k,l} F_{kl}} \tag{10}$$

The residues to be predicted to be in contact are the ones corresponding to the top F_{ij}^{APC}. This method does not predict all contacts, neither does always predicts correctly contacts, nonetheless is has been shown to be enough to assemble very precise 3D structural models [23].

Prediction of biological effect of mutations

Another important biological task that DCA is able to address with good results is the prediction of mutational effects in proteins. Protein mutational landscapes are mappings from nucleotide (or amino acid) sequences to phenotypes, quantifying therefore how mutations affect the biological functionality of proteins. Their comprehensive and accurate characterization is of central importance in medical biology: it can lead to the identification of genetic determinants of complex diseases [24] and it can guide our understanding of the functional contribution of genetic variations.

The score that can be obtained from DCA is given by the difference in energy between two sequences. It is used as a proxy for the effect of the mutation. In the case of a single mutation at site i:

$$\Delta E(a \to b) = H(a_1, ..., a_i = b, ..., a_N) - H(a_1, ..., a_i = a, ..., a_N) \tag{11}$$

is the score.

If this value is positive, it means that the mutation is potentially harmful. On the contrary a negative ΔE indicates a beneficial mutation. This is related to the form of the Boltzmann distribution of eq. (1) for which a lower energy means a more probable sequence. Interestingly the fitness computed this way is well correlated with different phenotypes, ranging from structural stability to antibiotic resistance. Remarkably DCA is able to get information about all those phenotypes and could even - in principle - have predictive power in assessing the effect of mutations that do hinder the ability of proteins to perform their physiological function per se .

Indeed, mutations might also be deleterious if they cause negative effects on one of the countless other cellular processes [25] and DCA could learn that relying solely on evolutionary data.

The scores obtained in this way can be compared versus fitness experiments that can score hundreds of single and double site mutants of a protein to empirically measure the biological effect of the mutations. As shown by Figliuzzi et al.[5] the score obtained in this way are much better predictors of the biological fitness of the mutations than obtained from independent (or equivalently $J_{ij} = 0$) models.

In general protein fitness landscapes are highly non trivial, this fact is well captured by the concept of epistasis: the context dependence of mutations in a protein. In presence of epistasis the effect of mutations is non additive, two single point mutations could be beneficial when considered alone, but could even become deleterious when present at the same time in a protein. This is one of the reasons why the DCA model performs well in this context, because positive and negative interaction between sites (encoded in positive or negative J's) can partially account for those complex interactions.

Artificial generation of homologous protein sequences

Another very interesting application of DCA is the ability to generate functional sequences. The task is far from being trivial, since sequence space for all possible proteins is enormous: for a 200-site protein the number is close to $20^200 = 10^{260}$. In this space only a very small fraction of sequences is functional, that is folds into a specific shape, and an even smaller number of them is part of a specific protein family, as currently the PFAM database contains more than 18 000 families.

Understanding if a specific amino acid sequence is a functional protein belonging to a specific family is therefore a very hard task. Nevertheless is was shown in [6] that by simply drawing sequences from the Boltzmann distribution is was possible to generate functional sequences, including a set of sequences with less than 65% identity to any of the other proteins in the natural MSA.

This result is highly non trivial and does not depend only on the fact that the Potts model fits one and two-site statistics. What is crucial here is the ability of to reproduce (without fitting) higher order statistical features from the natural MSA. In particular is was shown that the Potts model inferred by Boltzmann learning can capture 3-point correlations, the distribution of mutations in the protein family and the PCA representation of natural sequences [26].

2.4 Example: Beta-lactamases

Recently, two protein evolution experiments generated a great number of sequences belonging to the Beta-lactamase family of proteins (starting from the TEM-1 and PSE-1 sequences) with the aim to learn important structural and functional information from experimentally generated data instead of natural sequences MSA. We

are interested in understanding and modeling those experiments through the Potts model described above. To do so, we first check that we can infer a model for this protein family to extract relevant biological information, then we show that we can accurately sample sequence space to highlight the generative properties of the model.

Biological relevance

The Beta-lactamase family is a collection of bacterial enzymes that provide antibiotic resistance to the class of beta-lactam antibiotics, like penicillin, ampicillin or amoxycillin. Drugs belonging to this class constitute 60% of the worldwide antibiotic usage, and are among the most common and effective agents in the treatment of infectious diseases [27]. Resistance to beta-lactam antibiotics is gained by cutting their common molecular structure, the four atom ring known as I^2-lactam, to deactivate their antibacterial properties.

We are particularly interested in TEM-1, a particular Beta-lactamase found in *E.Coli*. The relative gene is located on a plasmid, thus is not present in every invidual of the specie, but is becoming more and more frequent over the years due to massive use of beta-lactam antibiotics.

Emergence of antibiotic resistant bacteria has been a natural outcome of the evolutionary process and strong selective pressure. The Center for Disease Control and Prevention states that antibiotic resistance is one of the biggest public health challenges of our time. Each year in the U.S., at least 2.8 million people get an antibiotic-resistant infection, and more than 35,000 people die [28].

Gaining structural and evolutionary knowledge about this protein family is therefore of primary relevance to address this increasing medical risk.

Features of the family and learning

To learn a Potts model on this protein family we built a MSA of natural sequences. To align the sequences we choose the PFAM Hidden Markov Model profile of the Beta-lactamase2 family (PF13354). We scan the NCBI RefSeq protein sequences database [29] for sequences belonging to the family. We filtered those sequences to keep only those not too gapped (at least 80% of coverage) and not too similar with each other (80% or less of sequence identity). We obtained a total number of 7515 sequences, which we consider as a high quality MSA of natural sequences. Once constructed the MSA we learned the Potts model of eq.(2) with a standard implementation described in [26].

A first view on natural sequence diversity

To analyze the natural sequences we start by looking at the distribution of the number of mutations among distinct sequences. To do so we employ the Hamming

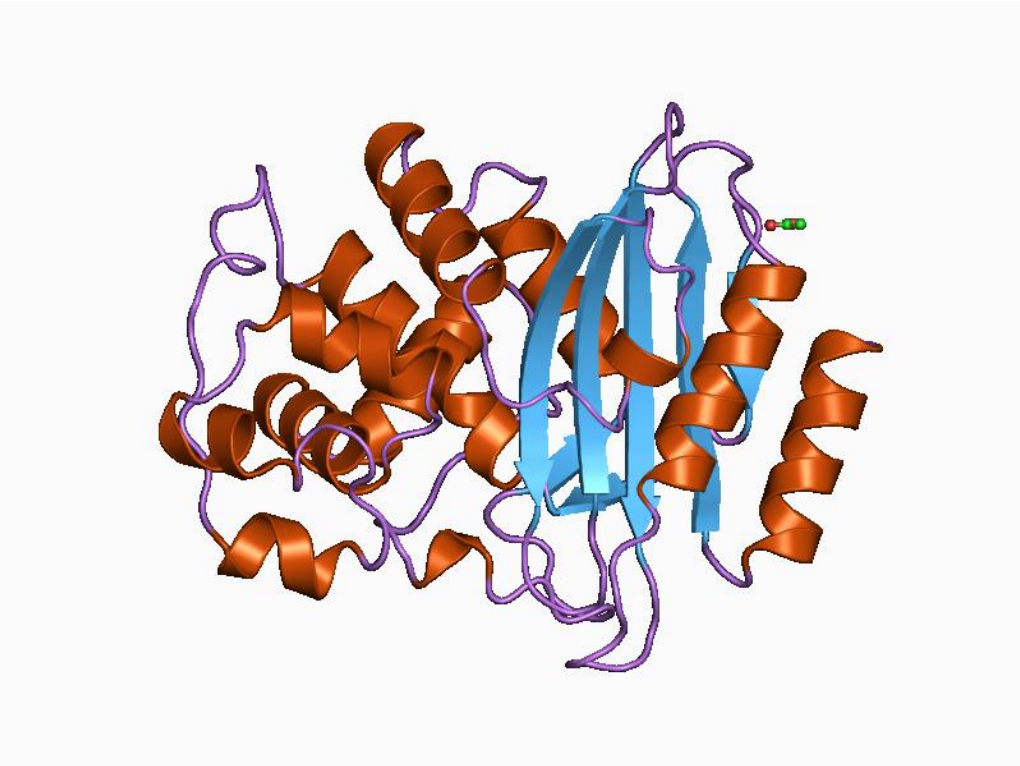


Figure 3: Three dimensional structure of the *Streptomyces albus* Beta-lactamase. Image by Jawahar Swaminathan and MSD staff at EBI

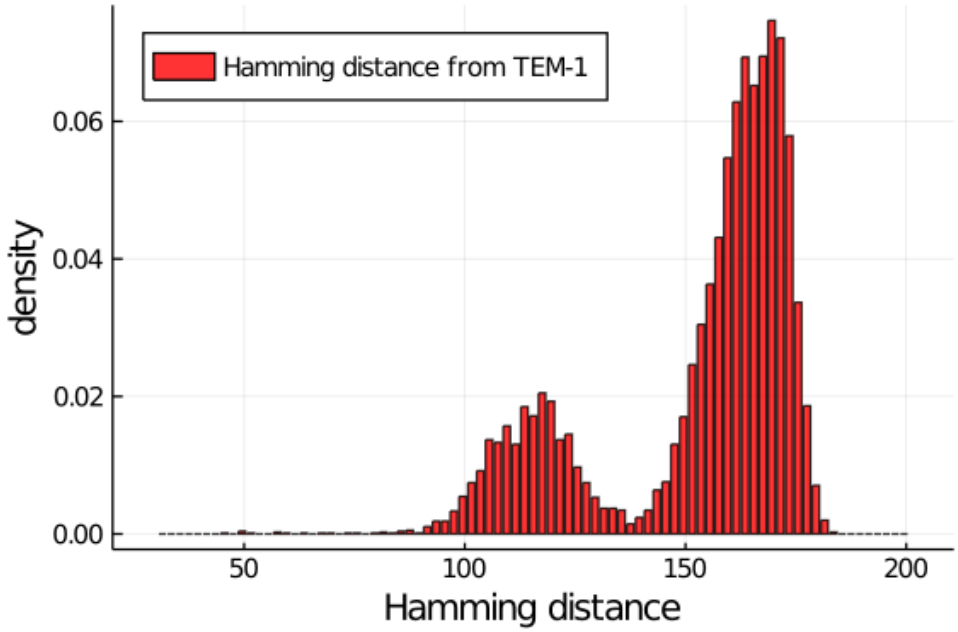


Figure 4: Histogram of the Hamming distances from TEM-1 of the sequences in our Beta-lactamase MSA.

distance. We chose TEM-1 as a reference and computed the Hamming distance of all sequences in the MSA with respect to it.

As we can see in Fig.5 two peaks are present in the distribution, the first one at $\sim 45\%$ identity with TEM-1 (Hamming distance ~ 110), the second at $\sim 15\%$ identity (Hamming distance ~ 170). The absence of sequences close to TEM-1 in the dataset depends upon an intentional decision made in constructing the MSA to avoid model overfitting and biases towards almost identical (i.e. non independent) sequences.

We also computed the pairwise distance between all sequences belonging to the MSA. To reduce the computational time and the memory requirements to perform the computation ($\propto \binom{M}{2}$) we sampled a subset of the MSA of $M' \sim \frac{M}{100}$ sequences and computed the distances between those sequences. Figure 5 confirms the great diversity in terms of amino acids of the sequences belonging to the dataset. Coherently with our construction of the MSA, no sequence is closer than 20%, i.e. ~ 40 amino acids in common, with any other sequence.

Principal Component Analysis

As we have seen from the previous histograms, we know that the family contains a very diverse set of sequences. A more graphical way to see the clustering of those proteins in sequence space is obtained by mean of Principal Component Analysis

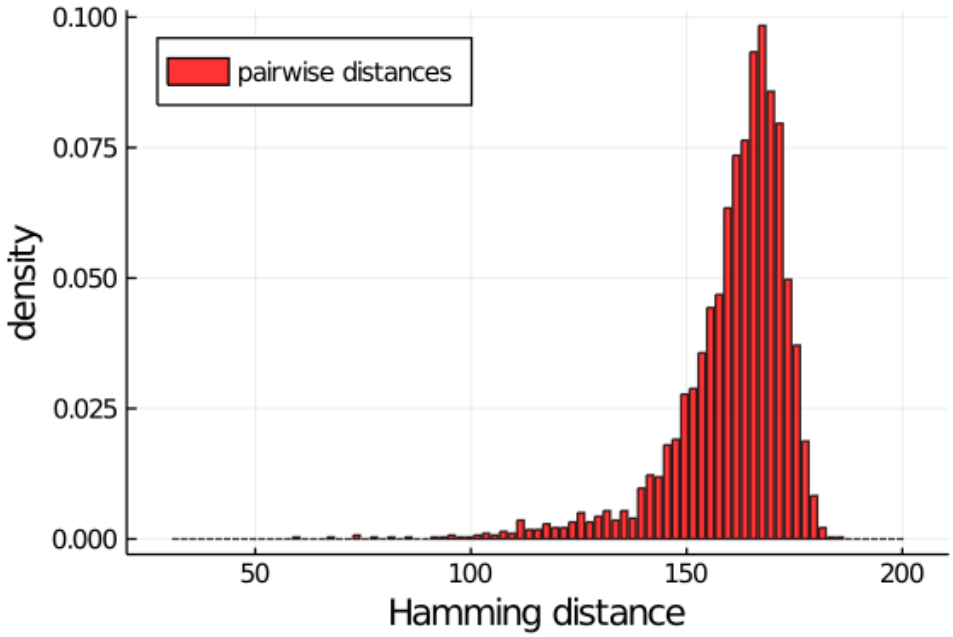


Figure 5: Histogram of all pairwise Hamming distances among the sequences of a subset ($\sim 1\%$) of our MSA.

(PCA). In Fig.6 we show all sequences mapped onto the first two components of the MSA.

We can recognise 3 main clusters. TEM-1, the sequence of our interest, is indicated in red and found in a smaller cluster. The sequences of this cluster correspond to the first peak in the distance distribution relative to TEM-1. The absence of the first peak in Fig.5 indicates that the main cluster in Fig.6 is very variable itself.

Prediction of biological effects of mutations

We then analyzed the sequences by mean of the Potts Hamiltonian described in eq. 2. We recall that the model assigns a probability to every possible amino acid sequence based on the likelihood of the sequence to belong to the family the model was trained on. An interesting application is given by the possibility of quantitatively scoring amino acid mutations. Thanks to next-generation sequencing the first experiments capable of experimentally assessing the effect of all single mutants of a protein are becoming more affordable. The kind of data produced is a formidable test set for any attempt to computationally assess the effect of mutations.

We consider the work of Ostermeier et al. [30]. They devised an experiment able to measure the difference in fitness given by single codon DNA mutants of the *E.Coli* TEM-1 Beta-lactamase gene and therefore of the TEM-1 protein. They generated a library of all single point amino acid mutations, expressed them in *E.Coli*, plated

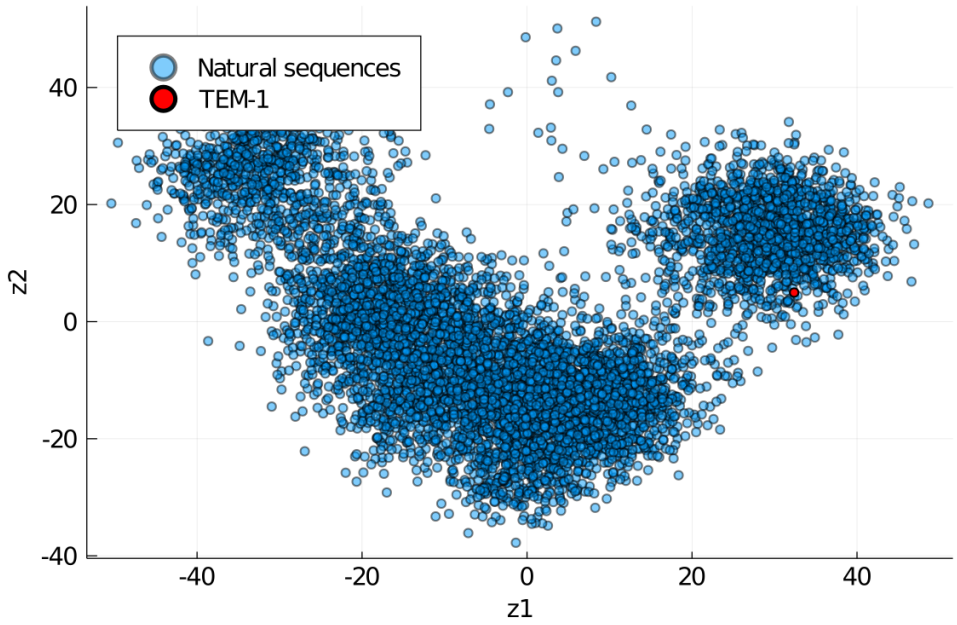


Figure 6: First two components in PCA space of all sequences belonging to the MSA of the Beta-lactamase family (natural sequences).

them in 13 dishes with different Ampicillin concentrations a_p with, $p = 1, .., 13$, and measured the fitness by counting the number of mutants after growth. They defined an unnormalized fitness f_i for every mutant i as

$$f_i = \frac{\sum_{p=1}^{13} c_{i,p} \log a_p}{\sum_{p=1}^{13} c_{i,p}} \tag{12}$$

with $c_{i,p}$ the count of mutants i after growth. They then normalized the value such that the fitness of the TEM-1 protein would be 1. As already mentioned the Potts Hamiltonian can be used to computationally predict the effect of mutations by mean of the difference in energy $\Delta H = H^{wt} - H^{muts}$. The prediction can be compared with the dataset just described.

We report in Fig.7 the scatterplot of the fitness score of all mutants generated by the experiment, compared with the difference in energies to TEM-1 of a computationally generated library of the same sequences. We note a good Pearson correlation between the two datasets (~ 0.70) and an even better Spearmann rank correlation (~ 0.74) , that is a measure that assesses how well the relationship between two variables can be described using a monotonic function, not just a linear one. This allows to take into account non linear effects in the relationship between our difference in energy and the fitness.

The majority of mutations with very low fitness (~ 0) are scored as deleterius by our model, that is they have an higher energy than the wild-type (TEM-1). We note

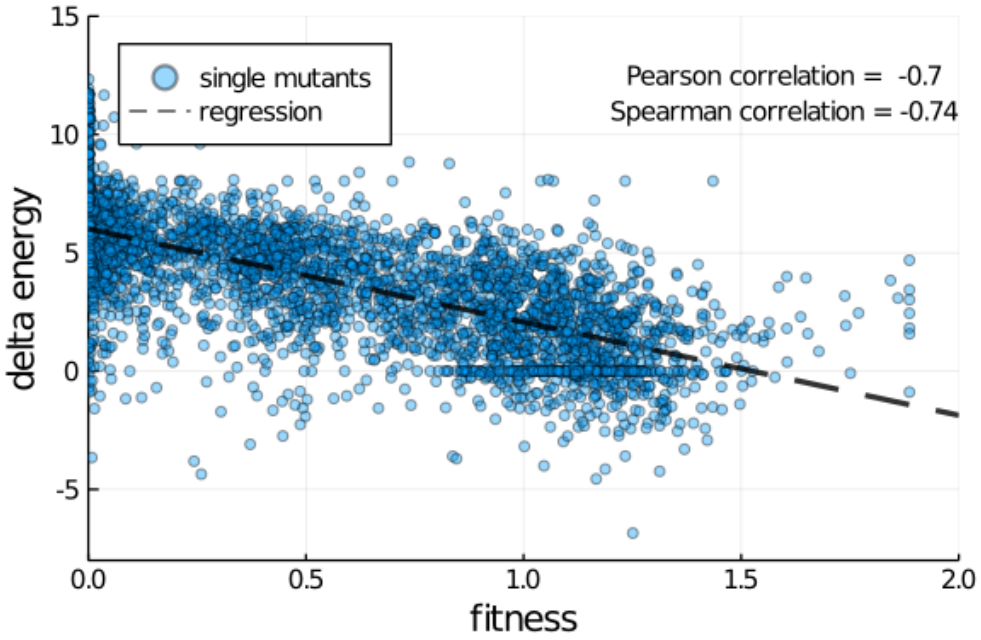


Figure 7: Scatterplot of the difference in fitness versus the difference in Potts energy of all single mutants of the TEM-1 protein sequence.

also a strip of points with 0 energy, but a relative wide range of fitness values centred around 1. This artifact can be explained considering that the experiment measured the fitness of all mutants, including the wildtype, for every site, thus resulting in multiple copies of TEM-1, that - due to experimental errors and possibly different codon transcription rates - had different fitnesses. The length of this strip can therefore be used as a proxy for experimental noise.

Energy of the sequences

The energy of the sequences can not only be used to asses the effect of single point mutations: the log probability of a sequence, up to an additive constant, can also be used as a global fitness score. Encouraging studies point in this direction: i.e. it was shown in [31] that the Potts energy could be used as a good predictor of folding in vitro for artificially generated sequences. We therefore decided to look at the distribution of energies of the sequences belonging to the MSA of natural sequences. We plotted the difference in energy of the sequences with respect to the reference (TEM-1):

$$\Delta H(\underline{a}) = H(\underline{a}) - H(\underline{a}^{ref}) \tag{13}$$

versus the Hamming distance of the sequences with respect to TEM-1. This analysis is motivated by the fact that one would expect that introducing random mutations to a given sequence would produce a protein unable to fold, or to be biologically

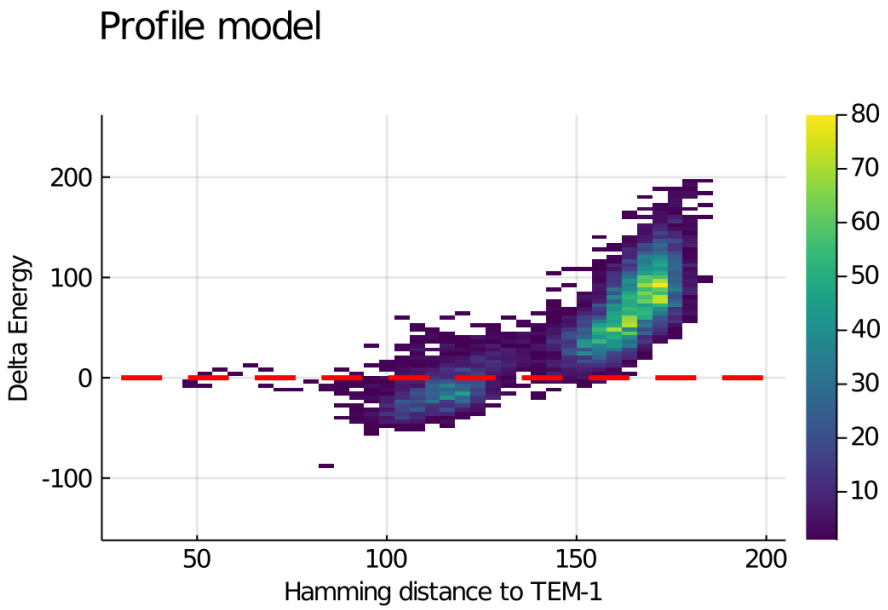


Figure 8: Difference in energy with respect to TEM-1 of the sequences belonging to our MSA of natural sequences, scored with the Potts Hamiltonian of eq.(2)

less functional. This would not be the case if the mutations preserved the 3D structure of the protein, its active sites, or in general if it does not effect the fitness of the host organism. This is the case for proteins belonging to the same family: a great sequence divergence still preserves function. A good sequence scoring function should therefore be able to predict a relatively limited fitness diversity between sequences belonging to the same family, regardless of their sequence divergence.

We start considering a simple sequence scoring function for sequences $\underline{a}$ based only on the frequency of amino acids $f_i(a_i)$ in the columns of the MSA constructed with the natural sequences of the family. This profile energy is therefore defined as:

$$H_{prof}(\underline{a}) = -\log P_{prof}(\underline{a}) = -\log\left(\prod_{i=1}^{N} f_i(a_i)\right) + C = -\sum_{i=1}^{N} \log f_i(a_i) + C \qquad (14)$$

This profile model does not include pairwise interactions between amino acids and can not capture epistatic effects between the sites. A plot of the difference in energies H_{prof} of the sequences belonging to the MSA of the Beta-lactamase family is shown in Fig.8.

We see from the plot that all sequences distant more than ~ 130 amino acids from TEM-1 score worse than TEM-1.

If we instead employ the Potts Hamiltonian as a scoring function a similar but richer picture emerges: some sequences belonging to the second peak now are also scored better than TEM-1.

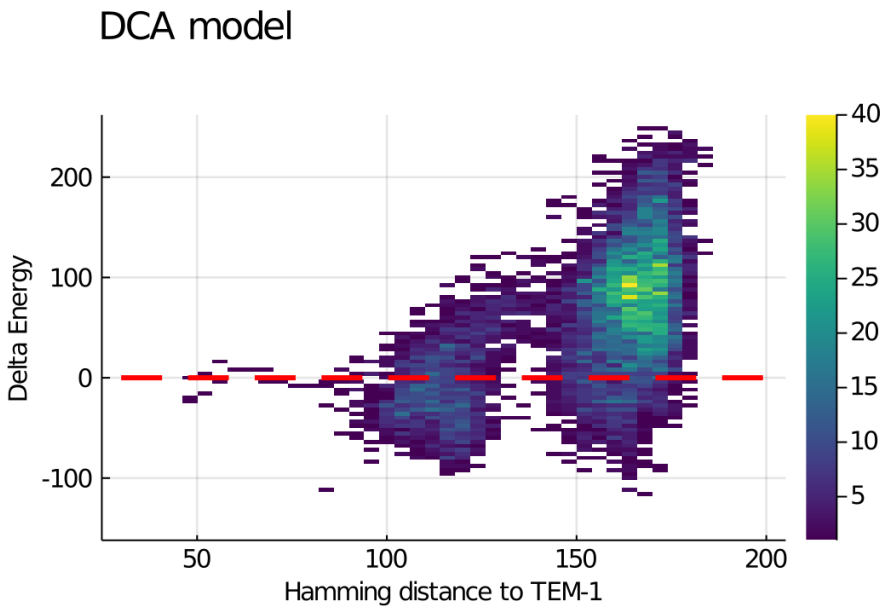


Figure 9: Difference in energy with respect to TEM-1 of the sequences belonging to our MSA of natural sequences, scored with the Potts Hamiltonian of eq.(2)

Crucially, however, both energy energy functions score sequences belonging to the protein Beta-lactamase protein family much better than randomized sequences at the same distance.

Contact map prediction

Probably the most successful application of DCA is residue-residue contact prediction. As already explained the couplings of the model J_{ij} can carry information on the physical proximity of the respective sites. To test the performance of the method, it is customary to compare the prediction to some known protein structure and to consider a Positive Predictive Value (PPV), that is the fraction of correctly predicted contacts (obtained from crystallographic experiments) out the total number of guesses of the model (that is the top scoring F_{ij}^{APC}).

Before going to the results for this protein family it is important to specify how do we define a contact. Two residue are in contact if any of the heavy atoms in the two amino acids are separated by less that 8Å in the folded state of the protein *and* the two amino acids are distant more than five residues in the linear polymer chain.

We report in Fig.10 the PPV curve for this protein family. The results are remarkable, showing a fraction of around 98% of correctly inferred contacts out of 100 predictions. We must also say that the results for this family are among the best across all protein families.

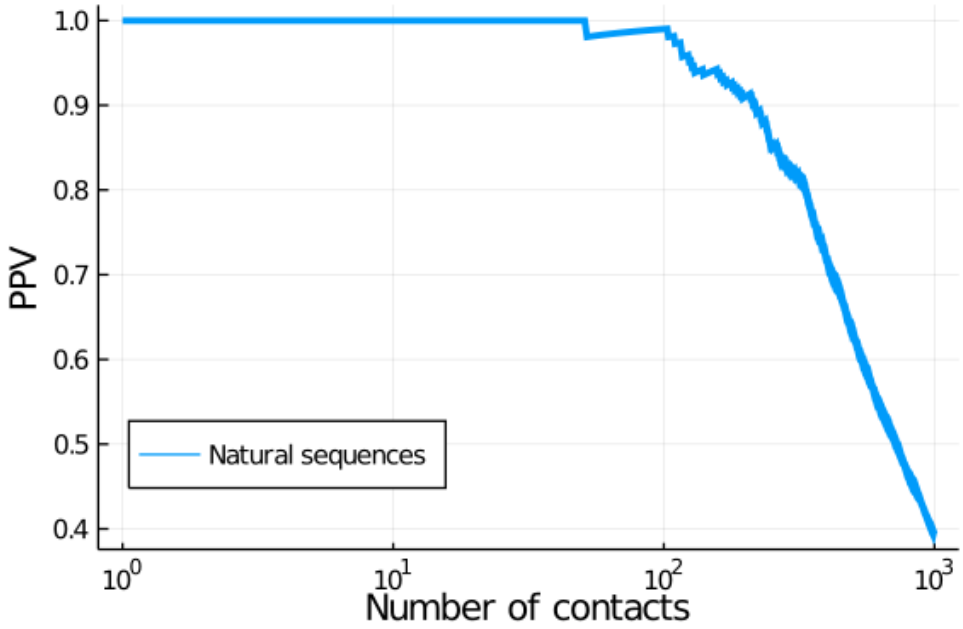


Figure 10: Fraction of the correctly predicted contacts for the Beta-lactamase family as function of the total number of predictions.

We also report a plot of the predicted contact map of the protein for the top $N = 202$ predicted contacts (Fig.11) and the underlying true crystallographic contacts. From the figure becomes immediately clear that the predicted contacts can be used for protein structure prediction thanks to their accuracy.

2.5 Generative properties of the model

We have described some quantitative features of the Beta-lactamase family of proteins and used the Potts model inferred by Boltzmann learning to predict biologically relevant quantities such mutational effects and residue-residue contacts. We now take a step further and we ask weather the model is generative, that is if it can statistically reproduce features of the MSA of natural sequences that were not fitted during the learning process. This would be a good hint in the direction of the DCA sequence landscape being generative. We could therefore employ it to study and simulate protein evolution experiments.

Equilibrium sampling from the model

To sample from the model we make use Monte Carlo Markov Chain (MCMC) sampling. This approach is needed to overcome the difficulty to sample from a model whose partition function Z is unknown and hard to estimate.

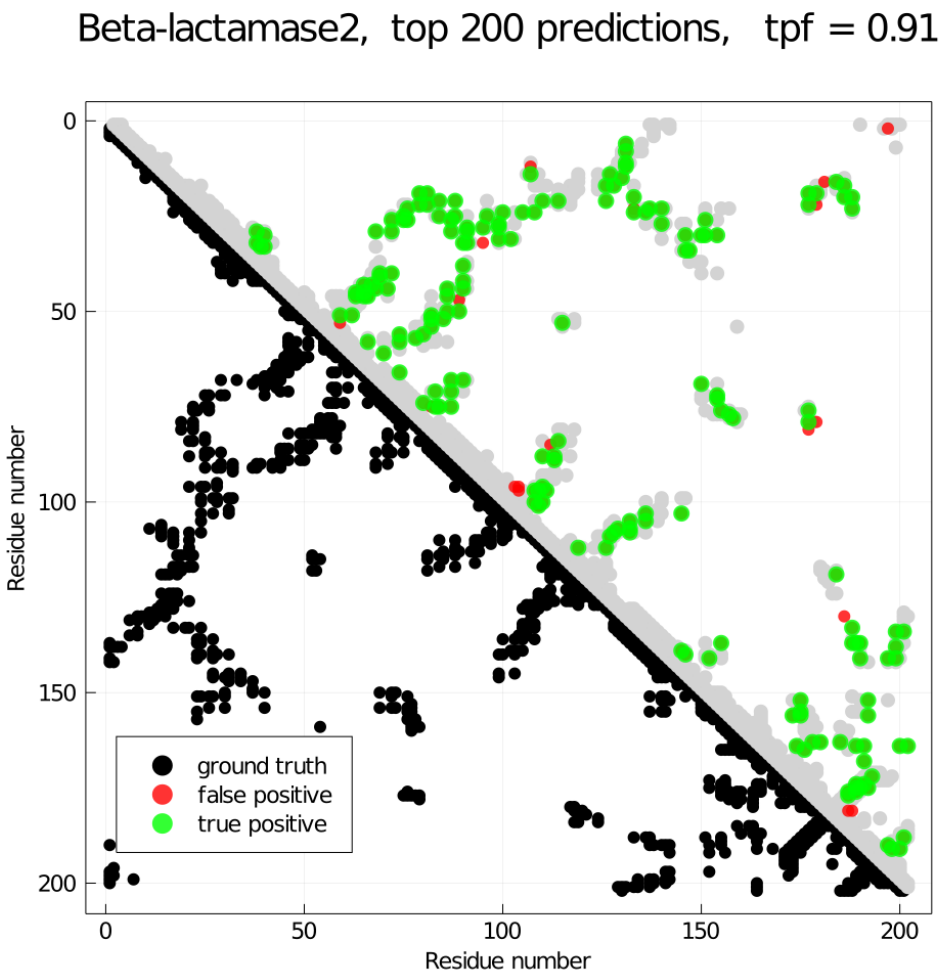


Figure 11: Contact map of the Beta-lactamase family. The dots represent contacts between the two sites indicated on the x and y axes. In black and grey (symmetrically) the true contacts extracted from crystallographic data, in green the correct DCA predictions and in red the false DCA prediction.

To this purpose we used Gibbs sampling, a special case of MCMC. This approach is useful when it is hard to sample from a joint probability distribution, but the conditional distribution of each variable are known and easy to sample from. This is exactly our case. The conditional probability $P(a_i = a|a_{k\neq i})$ can be easily computed:

$$P(a_i = a|a_{k\neq i}) = \frac{P(a_1, ..., a_N)}{P(a_1, ..., a_{i-1}, a_{i+1}, ..., a_N)} \propto \exp\left(h_i(a) + \sum_{\substack{k\leq j \\ (k \text{ or } j = i)}}^{N} J_{kj}(a_k, a_j)\right) \tag{15}$$

where the symbol $\propto$ indicates that the proportionality constant $Z^{-1}(a_1, ..., a_{i-1}, a_{i+1}, ..., a_N)$ does not depend on a_i. The value of the constant is easily found by summing over all amino acids in position i:

$$Z_i(a_1, ..., a_{i-1}, a_{i+1}, ..., a_N) = \sum_{b=1}^{21} \exp\left(h_i(b) + \sum_{k}^{N} J_{ki}(a_k, b)\right) \tag{16}$$

exploiting the property that $J_{ij}(a, b) = J_{ji}(b, a)$.

The sampling scheme is the following:

- Choose randomly a site i out of the $N = 202$ sites of the protein
- Compute $P(a_i = a|\underline{A}_{j\neq i})$ for every amino acid a, sample from this distribution and substitute the sampled amino acid with the previous one
- Repeat

If the Markov Chain is ergodic, the procedure is guaranteed to sample from the Boltzmann distribution: indeed Gibbs sampling satisfies the detailed balance condition.

Amino acid sequences obtained in this way are correlated. To get iid sequences we need to wait a number of steps greater than the equilibration or mixing time, that is finite for our probability distribution.

We have a way a to sample from our target distribution P, now we want to see how good the sampling is. To do so we concentrate on statistical quantities that we have not fitted during training. The analysis of the following sections are based on a MSA of 7515 sampled sequences.

Sequence distance distribution

The first quantity that we consider is the distribution of mutations. We consider again two different ways to test weather we can reproduce the statistical properties of the natural MSA in terms of mutations. The first one is the distribution of pairwise Hamming distances between the sequences in the sample. And as we see

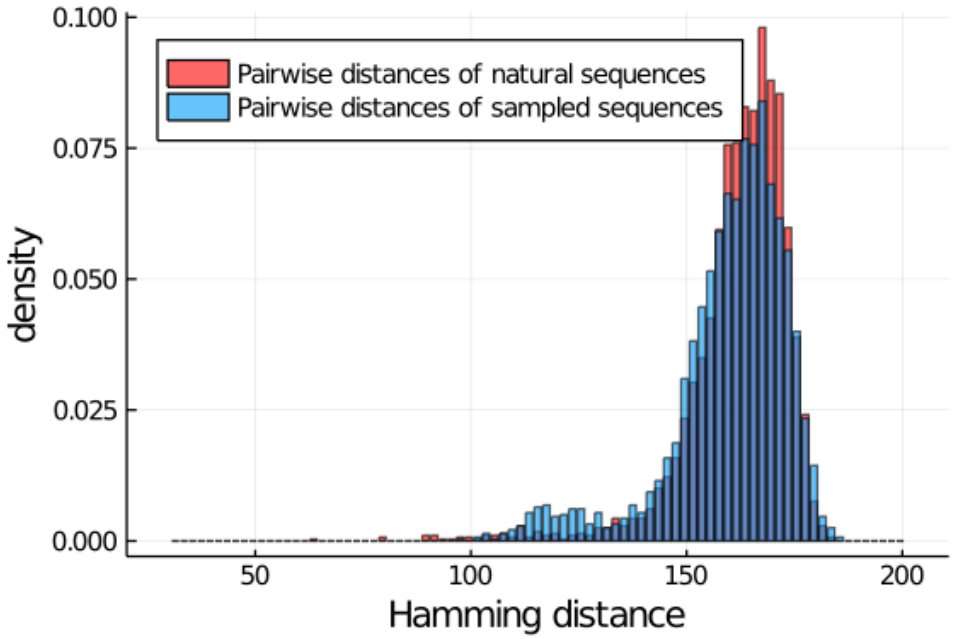


Figure 12: The two histograms show the distribution of pairwise Hamming distances among the sequences belonging to the natural and the sampled MSAs of the Beta-lactamase family.

in Fig.12 the sampled sequence resemble quite closely the distribution of pairwise mutations of the nautural ones.

The second test is to consider the distribution of mutations with respect to TEM-1. Again a visual inspection from Fig.13 shows that we reproduce the distribution of mutations of the natural sequences.

Principal Component Analysis

Secondly we control wether sampling from the model can reproduce the clustering of the sequences observed in PCA space. Again the result is very good, as we can see from the following figure.

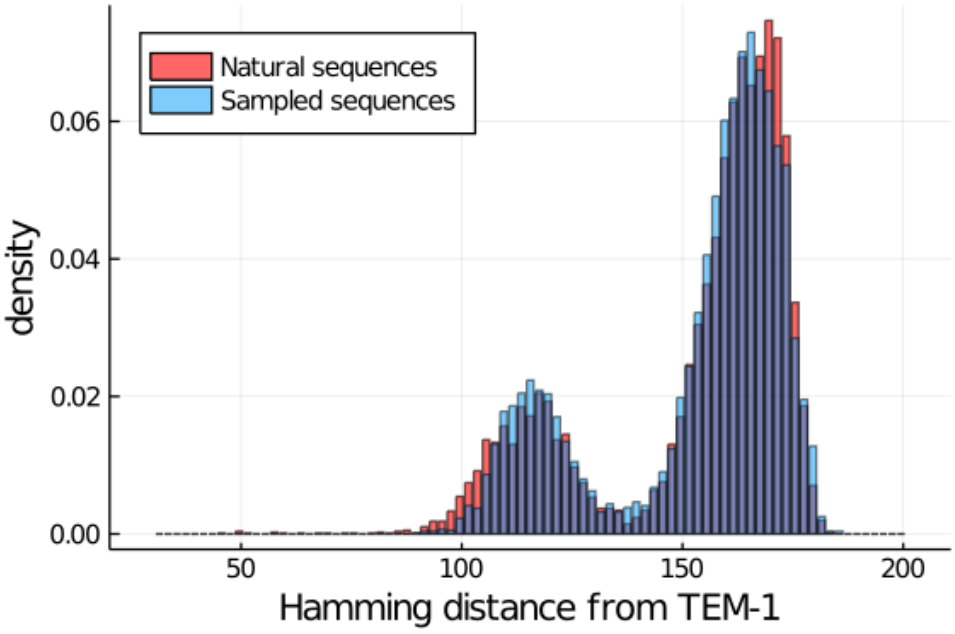


Figure 13: The two histograms show respectively the distribution of the Hamming distances from TEM-1 of the sequences belonging to the natural and the sampled MSA of the Beta-lactamase family.

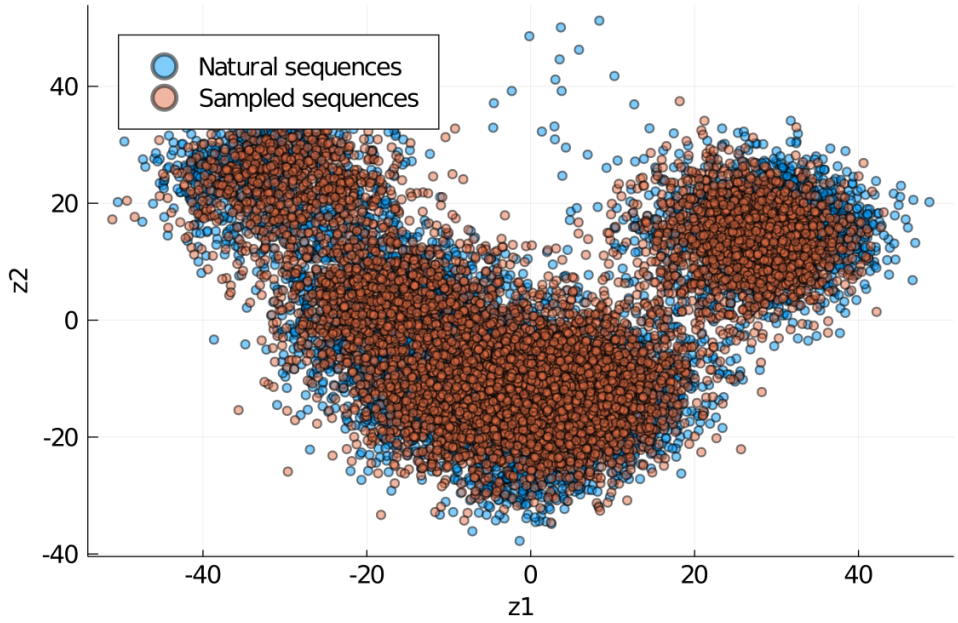


Figure 14: First two components in PCA space of all sequences belonging respectively to the MSAs of natural and sampled sequences.

3 In-vitro protein evolution

In this section we introduce the concept of laboratory protein evolution and its application to protein contacts prediction. We first discuss concepts of protein fitness landscapes and artificial protein evolution. We highlight the great developments and potential that lie beyond protein engineering applications of in-vitro evolution, worth the 2018 Nobel Prize for Chemistry to Frances H. Arnold "for the directed evolution of enzymes." Successively we present the results of two recent protein evolution experiments aimed at protein structure reconstruction via artificial sequence variation. We analyze them through the lens of DCA and we briefly discuss their differences. To write the introductory biological sections we took inspiration from the following papers: [32, 33, 34].

3.1 Natural evolution

Biological evolution can be formulated in terms of a dynamical process in a fitness landscape that acts at population level. It is based on two key principles: variation and selection. Mutations act as the basic working material to enable diversification. They occur in DNA molecules, in the form of nucleotide changes, through various chemical, mechanical and physical mechanisms. The most immediate effect of mutations at DNA level - in case they happen in a region coding for proteins - is the translation of such mutations at the amino acid level. This mechanism provides the necessary source of variation for selection to act upon. Evolution occurs when natural - or artificial - selection processes act on this variation, resulting in the spread or reduction of certain characteristics within a population in an heritable fashion. The spread of a specific trait - a phenotype - is related to its fitness, that is the ability of such phenotype to help its host survive in its environment.

Proteins are one of the simplest and best examples of evolvable biological systems. Just few mutations are enough to alter their biological function and yet they are quite mutationally robust; an example is provided by the Deep Mutational Scanning data of TEM-1: some mutations do not change much the host fitness, whereas others can have dramatic effects.

3.2 Protein fitness landscapes

Protein evolution can be formulated in terms of protein sequence dynamics in a fitness landscape. In such landscape a fitness function assigns a value to each protein sequence pushing sequences to move towards areas of higher fitness over time. Fitness landscapes are intrinsically high-dimensional and complex. Given a protein of length N, it can be embedded in a N-dimensional amino acid categorical space, where each point of the space corresponds to one of the 20^N possible amino acid sequences the protein can assume. This space is the domain of the fitness function, itself a potentially complicated function of the protein phenotypes.

With this picture in mind protein evolution can be viewed as the dynamical jumping from one functional protein to another in the amino acid sequence space. Clearly nature has not sampled randomly this space - during a billions year long struggle for survival- but has followed the routes set by the landscape.

3.3 Experimental protein evolution

Notwithstanding significant advances in the field of structural biology and the recent sequencing revolution, a molecular-level understanding of the individual differences of protein function remains elusive. Equally, maps from amino acid sequences to phenotype and function are expensive and difficult to obtain and predicting the amino acid changes able to generate a specific behavior remains a challenge. However, we know that evolution was able to generate thousand of proteins with impressive and very specific functions.

Laboratory protein evolution refers as a general term used to describe various techniques for generating protein mutants and selecting them according to a specific phenotype, typically repeating this process over multiple rounds, mimicking natural evolution.

Performing protein evolution in the lab has proven a powerful tool to investigate natural evolution [35], to address central questions in the biophysics of proteins [32] and as a general purpose engineering principle to evolve new biochemical functions [36].

A specific application is directed evolution: a laboratory tool for optimizing protein function to produce new, non-natural tasks. To understand it we can employ again the analogy of evolution as a walk on this high-dimensional fitness landscape. If proteins can be efficiently selected according to a desired - not naturally occurring - function, then regions of higher fitness represent target proteins, and iterations of mutation and artificial phenotypic selection explore the space and include new sequences while going uphill.

3.4 Contact prediction via in-vitro protein evolution

In recent works [1, 2] the authors wondered whether the tools of DCA could be applied not only to natural sequences, but also to other types of evolutionary data, such as ensemble of proteins evolved in vitro. Building upon the ability of DCA to obtain structural information of proteins, this approach could provide structural information for proteins difficult to crystallize or simulate, such as disordered or membrane proteins.

One of the other advantages of an evolution-based approach relying on artificial data would be the possibility to obtain sequence data for proteins lacking large databases of homologous sequences and to be able to increase (almost) at will the amount of data used to train the algorithms.

The two experiments employed evolution to generate a collection of functional variants of two proteins belonging to the Beta-lactamase2 family by expressing them in *E.Coli* and coupling a targeted mutagenesis of the gene to a selection pressure for antibiotic resistance.

Fantini's experiment

We briefly report the experimental procedure employed in the article of Fantini et. al. [1] and then we analyze more in depth the sequence data produced by their experiment using the tools of DCA.

The authors employed error-prone PCR (epPCR) [37] to generate a large library of variants of the TEM-1 protein, followed by transformation into bacterial cells and in vivo phenotypic selection with ampicillin at $25\mu g/ml$. The plasmid library carrying the mutants was then collected from bacteria that survived selection and sequenced.

To be able to produce a great amount of bacterial colonies and to conserve the complexity of the libraries, the colonies were incorporated in a semisolid medium, exploiting the advantages of both solid and liquid cultures. After colonial growth the plasmid library were collected from the media by centrifugation. In total 12 generations of mutation and selection were performed. The 1-st, 5-th and 12-th generations were sequenced with the Pacific Bioscience (PacBio) Sequel platform and analyzed. The process is schematically depicted in Fig.15.

The group was able to obtain one of the most diversified libraries of laboratory evolved proteins ($\sim 10\%$ of sequence diversity) with around $100k$ sequences in the 5-th and 12-th generation. The sequences collected in MSAs were used to learn a Potts model with a standard implementation of plmDCA [18] and to extract the top-scoring predicted residue-residue contacts. The predicted contact map partially matched that of the reference crystal structure with a bias toward short and medium-range contacts.

Sander's experiment

We describe also the experimental details and key results of the second protein evolution experiment from Sander et.al. [2]. The group subjected two bacterial antibiotic resistance proteins - the Pseudomonas Beta-lactamase PSE-1 and aminoglycoside acetyltransferase AAC6 — to experimental evolution by repeated rounds of mutation and selection for preservation of function. We will concentrate in the following only the PSE-1 experiment, since it belongs to the Beta-lactamase family.

To promote sequence divergence, they applied a high mutation rate using ep-PCR and they selected for functional proteins under permissive selective conditions (6 mg/mL ampicillin). The antibiotic concentration was slightly above the minimal inhibitory concentration, MIC, for *E.Coli* lacking a resistance gene. Successive rounds of mutation and selection were applied by using the selected sequences in one

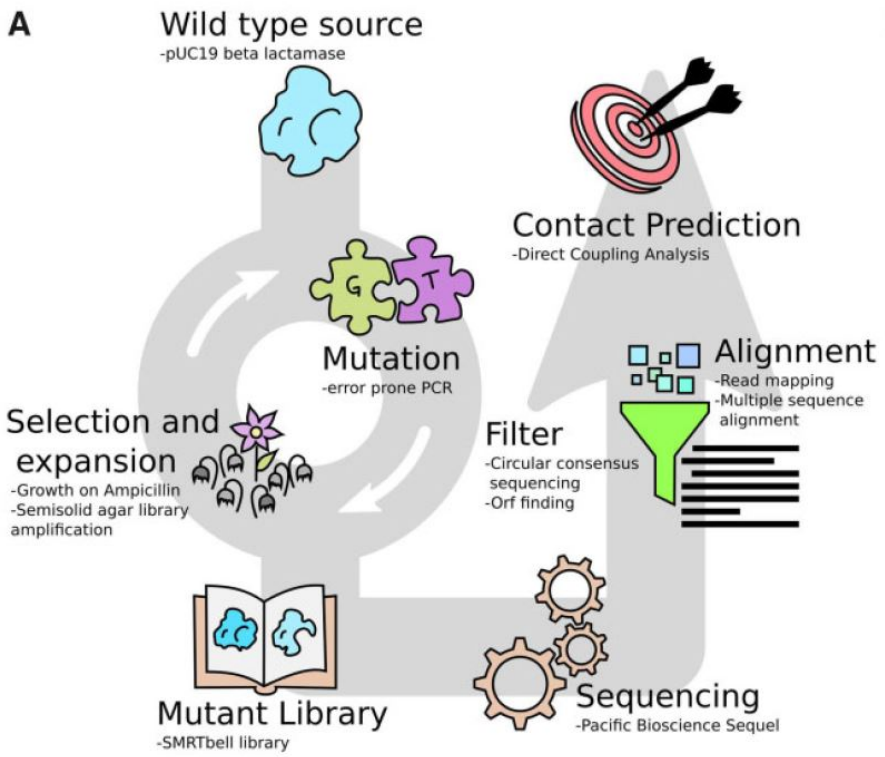


Figure 15: Schematic representation of the protein evolution protocol implemented by Fantini and collaborators. Source: [1].

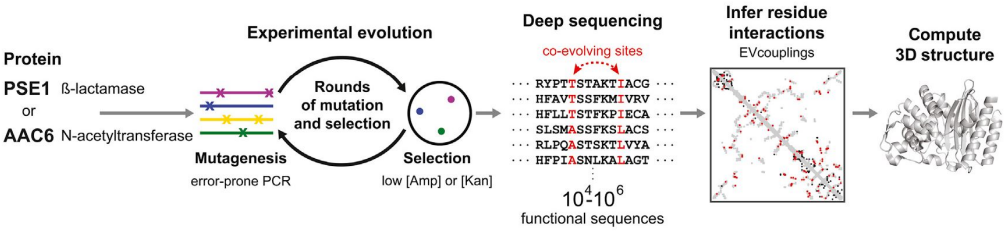


Figure 16: Schematic representation of the procedural steps leading to protein structure prediction from experimental protein evolution. Source: [2].

round as the template for mutations in the next round. The sequences belonging to rounds 10 and 20 were sequenced and collected in two MSA. By applying DCA to the artificially generated sequences the authors were able to successfully compute the 3D structure of the protein.

3.5 A visual tour of Fantini's experiment

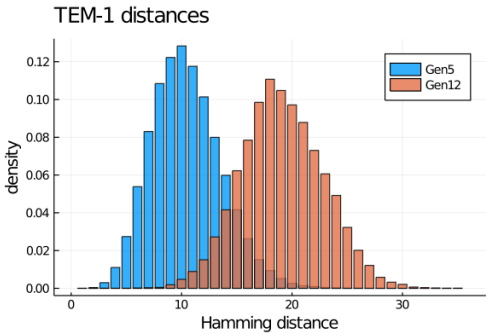
Before modeling protein evolution experiments in the DCA energy landscape, we analyse variability and energies of the experimentally generated sequences. This will serve as a test for the quantitative character of our modeling approaches, which are presented in the following chapter.

We analyze here the sequence data coming from the 5th and the 12th generation of the Fantini's experiment. We first collected the sequences in two MSAs by aligning the them to the PFAM Hidden Markov Model of the Beta-lactamase2 protein family. Then we processed the sequences to obtain two clean MSA. We removed all sequences with more than 6 gaps as well as sequences with detectable alignment errors. We obtained two MSA for the 5th and the 12th generation containing respectively 99201 and 34431 unique sequences.

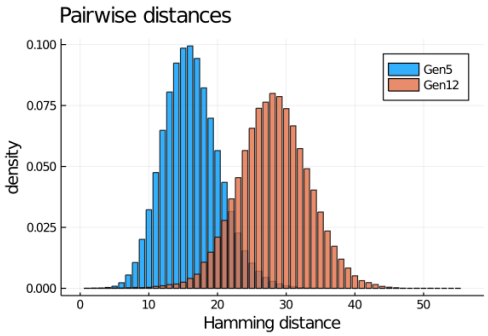
Sequence distance distribution

The experiment was able to generate a great amount of sequences, nonetheless they had much less sequence diversity than the dataset of natural sequences. We can see from Fig.17b that the mean sequence identity between any two sequences in the two datasets is ~ 15 and ~ 28 for the two generations.

Looking at Fig.17a we see that the 5th generation was less diverged with respect to the wild-type TEM-1 than the subsequent generation 12, as expected by a generational experiment. The majority of the sequences from the two generations diverged less than 15% (30 mutations) from TEM-1.



(a) Hamming distances from TEM-1 of all sequences belonging to the 5 th and 12th generation of Fantini's experiment.



(b) Pairwise Hamming distances of a subset of the sequences belonging to the 5 th and 12th generation of Fantini's experiment.

Figure 17

Principal component analysis

We plotted the first two components obtained by PCA of the two MSAs to visualize the clustered organization of the sequences in PCA space. We can see in Fig.18b that the sequences from the two generations occupy rather uniformly PCA space around TEM-1. As expected by the low sequence divergence of the experimental sequences compared to the natural ones, they span a smaller subset of the PCA space, clustering around TEM-1.

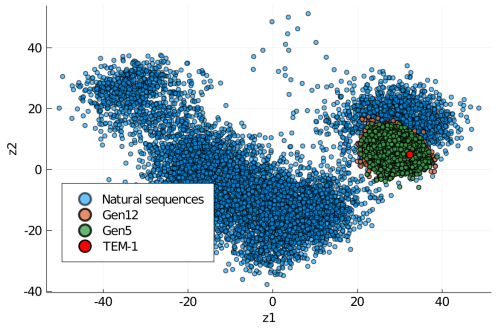
Site entropy

As we already know, the amino acids in the columns of the MSA of natural sequence are not equally conserved among all sites. As a measure of the site amino acid variability for the two generations we employed the Shannon entropy. Given a site i, the site dependent entropy S_i is defined as

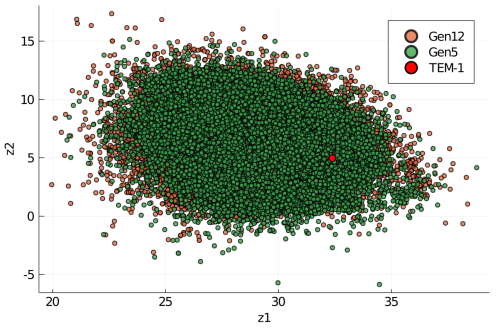
$$S_i = -\sum_{a=1}^{21} f_i(a) \log(f_i(a)) \tag{17}$$

where $f_i(a)$ is the frequency of amino acid a in the column i of the MSA of the sequences belonging to the relative generation.

We report in Fig.19 the site entropies of the two generations. They present a great degree of correlation (Pearson correlation ~ 0.9). As expected the variability is less emphasized for the sequences belonging to generation 5, given that they are closer to each other and have diverged less from TEM-1. We recall that if the amino acids were uniformly distributed in a site i, its entropy would be $S_i = \log(21) \sim 3$.



(a) First two components in PCA space of the sequences generated by the 5th and 12th generation of Fantini's experiment. The natural sequences are plotted for comparison.



(b) First two components in PCA space of the sequences generated by the 5th and 12th generation of Fantini's experiment.

Figure 18

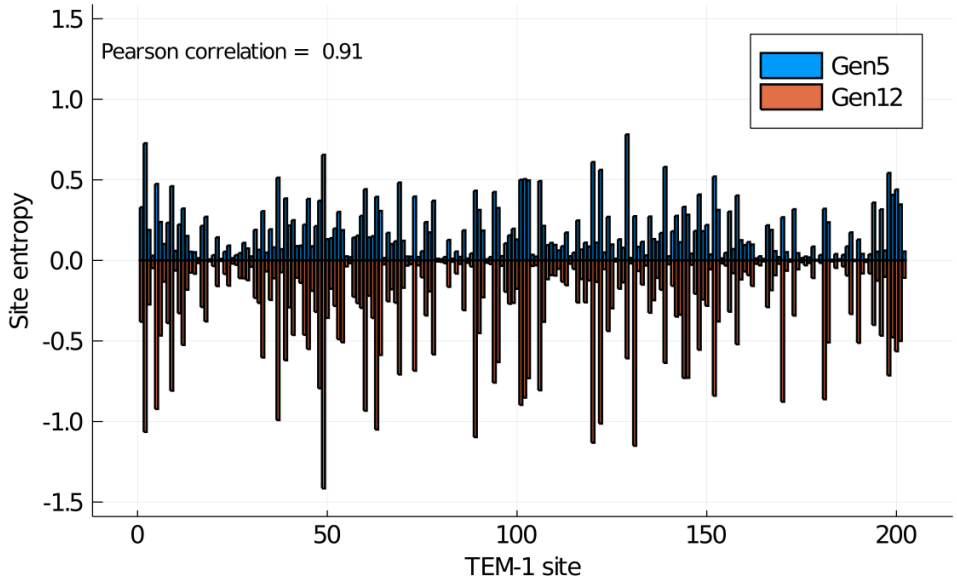


Figure 19: Shannon entropies of the distribution of the amino acids for every site of the MSA of generation 5 and 12 of Fantini's experiment. The Pearson correlation refers to the site entropies between the two generations.

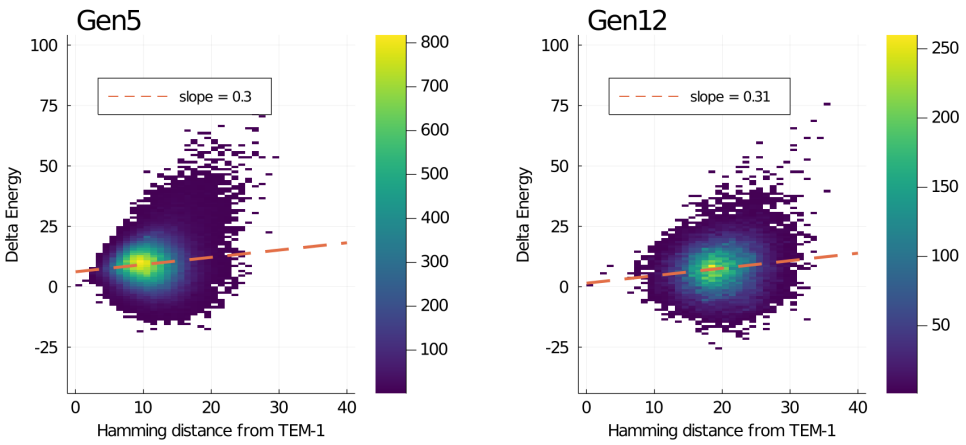


Figure 20: Density plot of the difference in energy versus the Hamming distance of the sequences belonging to the 5th and 12th experimental generation of Fantini's experiment. The color is an indicator of the number of sequences present in the spot.

Energy of the sequences

We turn now to the analysis of the energies of the sequences. We rely to the visualisation of difference in energy with respect to TEM-1, versus Hamming distance.

Fig.20 reports the energies for generation 5 and 12. The plot is very interesting and informative. Thanks to the experiment we know that the sequences belonging to the two generations have accumulated mutations, yet they have been selected for their beta-lactam antibiotic resistance capabilities, retaining their function. Coherently our model scores well the sequences and computes energies quite uniformly distributed around the energy of TEM-1, even for the most diverged sequences. Apart for a slight slope of the energy relative to the number of mutations, the former is almost independent from the Hamming distance to TEM-1. The results confirms the potential for the Potts energy to be used as a fitness scoring function for protein sequences, at least for this protein family.

Contact prediction

Finally we report the contact map prediction (Fig.21) for residue-residue contacts obtained by Fantini and collaborators by training a DCA model on the sequences of the 12th generation alone. The contacts are too sparse to define clear interaction zones and they tend to cluster around the diagonal. The strongest predictions from these mutational data concerns adjacent secondary structure elements.

Despite this partial results for structure prediction, the experiment was able to simulate the course of evolution by in vitro mutagenesis and selection generating an extremely valuable dataset of protein sequences.

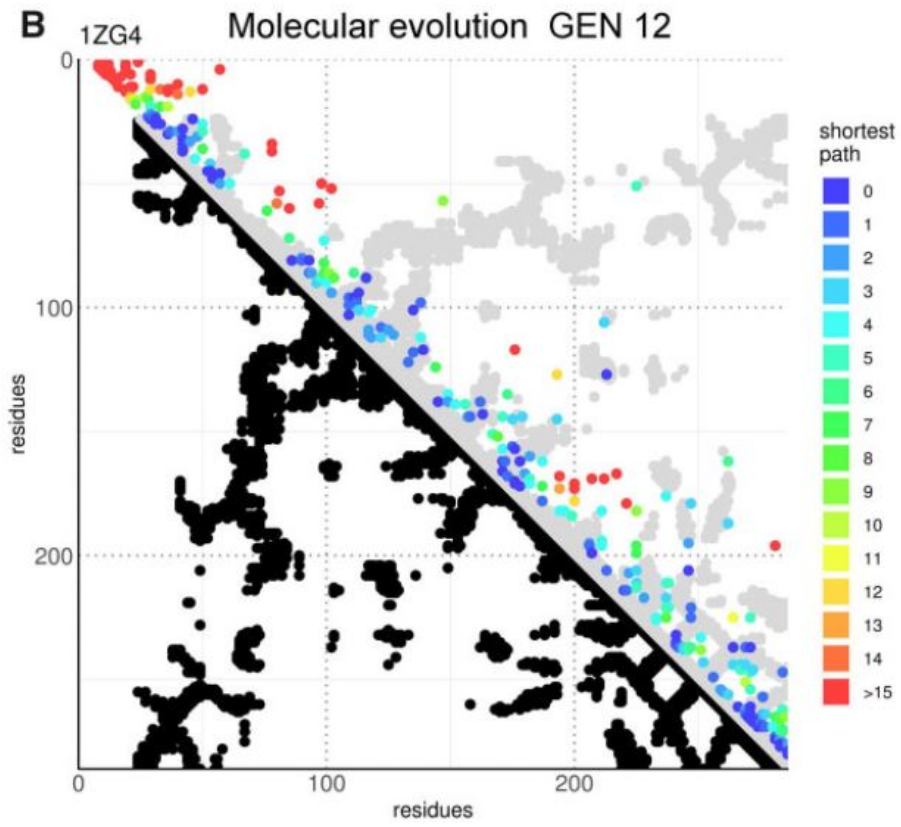
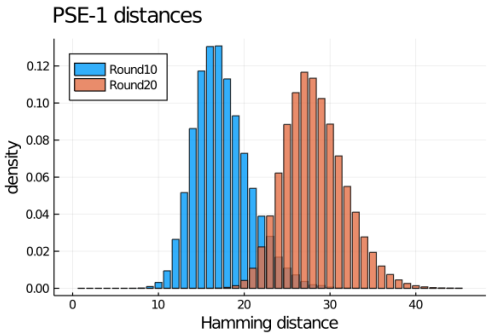
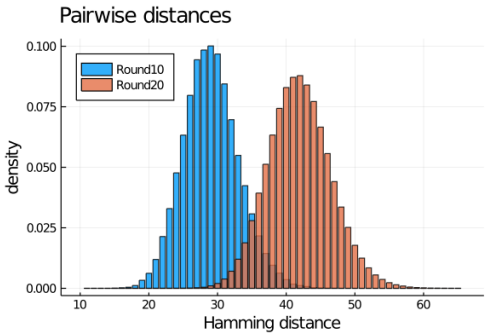


Figure 21: Contact map prediction from sequences belonging to the 12th generation of Fantini's experiment. Source: [1].



(a) Hamming distances from PSE-1 of all sequences belonging to the 5 th and 12th generation of Sander's experiment.



(b) Pairwise Hamming distances of a subset of the sequences belonging to the 10th and 20th rounds of Sander's experiment.

Figure 22

Differences with Sander's experiment

We briefly report on some differences between the two experiments that we presented. Compared to the experiment of Fantini, the Sander group employed a similar protein evolution procedure, though starting from a different protein, PSE-1. This protein sequence has an substantial (~ 100) sequence divergence from TEM-1. The MSAs coming from the 10th and 20th rounds of evolution contained respectively 1 and 162163 sequences after aligning and cleaning them in the same way described above. The sequences had greater sequence divergence from the wild-type compared to those of Fantini, as well as more inter-protein distance (cfr. Fig.22).

By looking at sequences in PCA space (Fig.23) we see that they span a slightly bigger area than those of Fantini, though staying in the same cluster.

Also in this case our energy function scores well the sequences. The result again is non trivial: the sequence landscape of PSE-1 is in principle very different from that of TEM-1, still we see similar features in the energy versus Hamming distance distribution. Worth noting here is the bigger slope (~ 1) of the energies. We speculate that this feature is related to a less stringent selection acted upon the sequences, allowing slightly deleterious mutations to accumulate. This could be one of the reason that allowed the group to predict with reasonable precision the protein contact map and then assemble a good structural model.

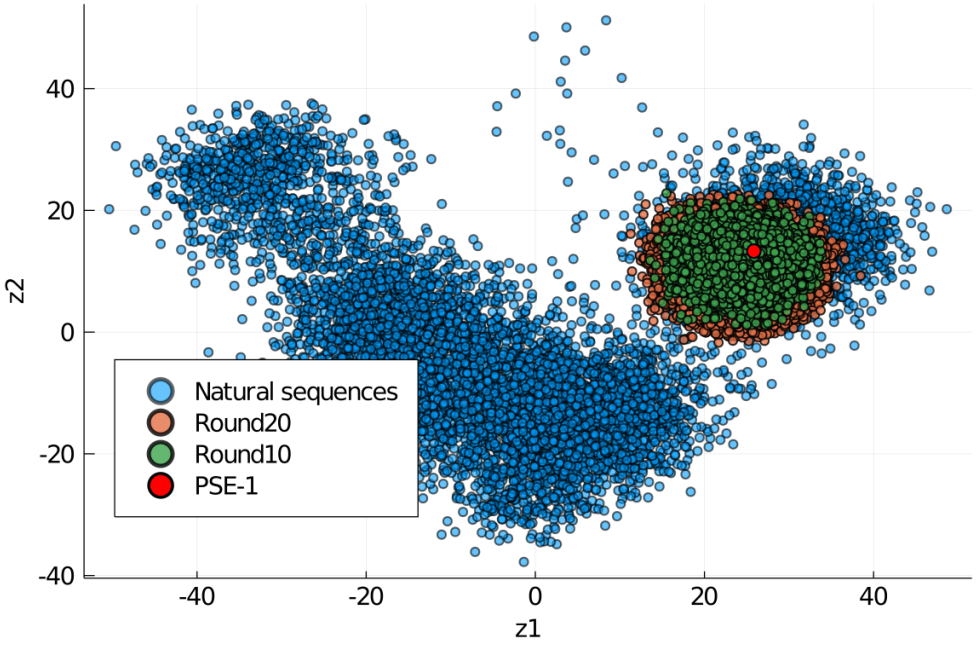


Figure 23: First two components in PCA space of the sequences generated by the 10th and 20th round of Sander's experiment.

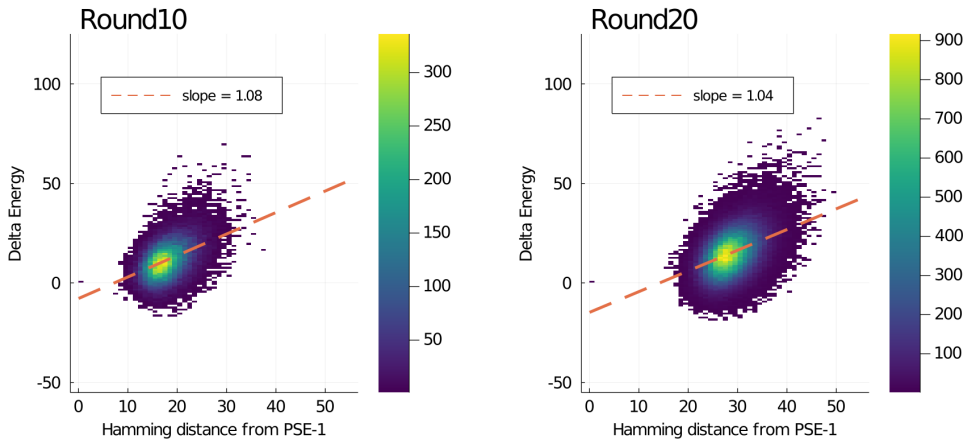


Figure 24: Density plot of the difference in energy versus the Hamming distance of the sequences belonging to the 10th and 20th experimental round of Sander's experiment. The color indicates the number of sequences present in the spot.

4 In-silico modeling of experimental protein evolution

In the previous sections we have presented the Direct Coupling Analysis method, its application to residue-residue contact prediction and mutational fitness prediction. We have demonstrated the generative properties of the model and we have also shown how sequence data coming from evolution experiments can be well described by the DCA sequence landscape. We have also understood how coupling protein evolution experiments with DCA could pave the way towards structure determination by artificial selection in vitro.

We now take a step further and we ask weather we can use DCA to simulate in silico protein evolution. To test our results we will check if experimental sequences are coherent with our simulations. Being able to simulate protein evolution experiments could allow researchers to tune a priori characteristics like the number of mutations per sequence, the strength of selection and the number of sequences needed to successfully determine protein structure from artificial experiments data.

4.1 Landscape sampling

To simulate protein evolution we rely on sequence landscape sampling. We start with a single sequence - that we will refer to as the "wild-type" - we choose randomly a site in its amino acid chain (to imitate ep-PCR) and we emit a new amino acid for the site according to a probability distribution on the 20 amino acids (to imitate selection). We repeat this process starting with the newly generated sequence to evolve the protein for a number of steps. In this way we have simulated the evolution of a single protein. To simulate a full protein evolution experiment we repeat the protocol presented above until we obtain a full library of evolved sequences that we save in the form of a MSA.

For the probability of emission of a new amino acid a we use the conditional probability $P(a_i = a|a_{j\neq i})$, with

$$P(\underline{a}) = \frac{e^{-H(\underline{a})}}{Z} \tag{18}$$

in this way we implement a Gibbs sampling procedure that locally samples the protein landscape defined by the Hamiltonian $H(\underline{a})$.

Our best candidate for $H(\underline{a})$ is the Potts Hamiltonian inferred from the dataset of natural sequences belonging to the family of the wild-type starting sequence. However, to prove this point, we consider other two possible choices for the Hamiltonian

Random sampling

The first Hamiltonian that we consider is a trivial one:

$$H_{rand}(\underline{a}) = 0 \tag{19}$$

The landscape described by this Hamiltonian is flat and every sequence has the same probability. As a consequence:

$$P_{rand}(a_i = a|\underline{a}_{j\neq i}) = \frac{P_{rand}(\underline{a})}{P_{rand}(\underline{a}_{j\neq i})} = \frac{20^{-202}}{20^{-201}} = \frac{1}{20} \tag{20}$$

that is every amino acid is randomly emitted independently from the residue position.

Profile sampling

The second Hamiltonian that we consider describes a so called profile model:

$$H_{prof}(\underline{a}) = -\sum_{i=1}^{N} h_i^{'}(a_i) \tag{21}$$

The fields $h_i^{'}(a_i)$, not to be confused with $h_i(a_i)$ of the Potts model, can be easily obtained from Maximum Entropy modelling, by imposing the marginals of the probability distribution to be equal to $f_i(a_i)$. One obtains:

$$h_i^{'}(a_i) = \log(f_i(a_i)) + C \qquad \text{with} \qquad C = \log\left(\sum_{b=1}^{21} e^{h_i^{'}(b)}\right) \tag{22}$$

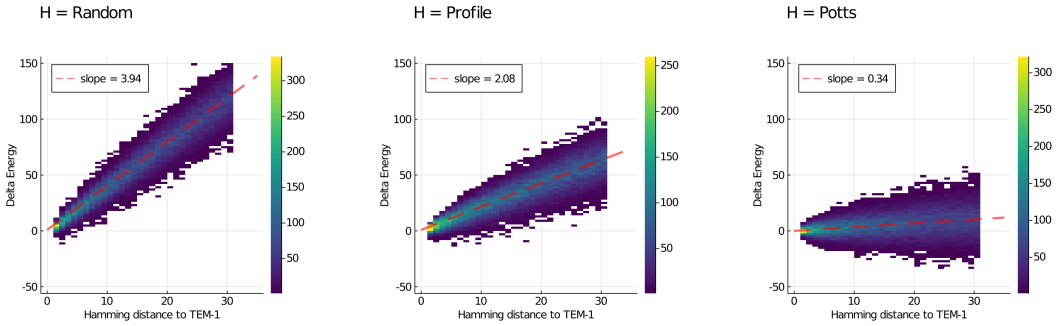
The landscape described by this Hamiltonian is not flat and already encodes some structure constraints. The emission probability is

$$P_{prof}(a_i = a|\underline{a}_{j\neq i}) = \frac{P_{prof}(\underline{a})}{P_{prof}(\underline{a}_{j\neq i})} \propto e^{h_i^{'}(a)} \tag{23}$$

were again it is meant that the proportionality constant $Z_i^{'-1}$ does not depend upon amino acid a. Its value can easily computed by normalisation

$$Z^{'} = \sum_{b=1}^{21} e^{h_i^{'}(b)} \qquad \text{leading to} \qquad P_{prof}(a_i = a|\underline{a}_{j\neq i}) = f_i(a_i) \tag{24}$$

that is every amino acid is emitted with a probability that only depends on the site and is equal to the re-weighted frequency of that amino acid in the natural sequences MSA.



(a) Random sampling. **(b)** Profile sampling **(c)** Potts sampling

Figure 25: Comparison of three sampling methods.

Potts sampling

The Hamiltonian here is the Potts Hamiltonian already introduced at the beginning:

$$H(a_1, ..., a_N) = -\sum_{i}^{N} h_i(a_i) - \sum_{i \leq j}^{N} J_{ij}(a_i, a_j) \tag{25}$$

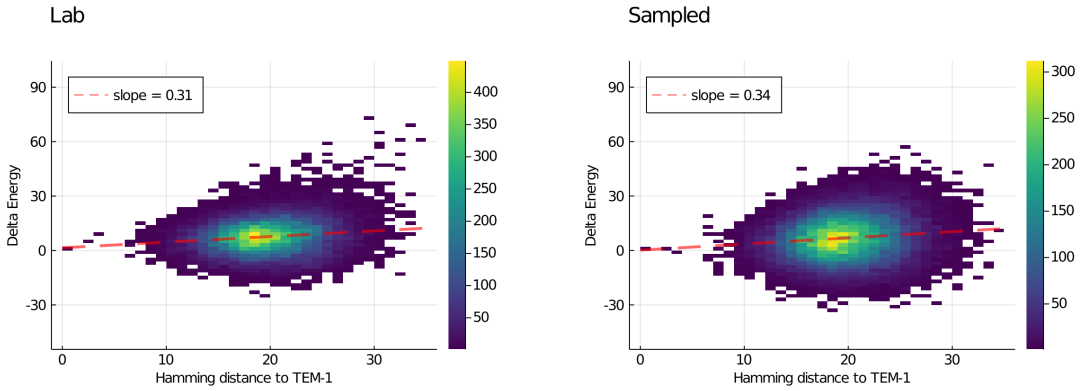
As we have seen the probability of emission of an amino acid in this case is context dependent, that is depends not only on the site, but also on the amino acids present in the rest of the chain.

To compare the sampling obtained by those three Hamiltonians we plotted the change in energy versus the Hamming distance to TEM-1 of three datasets of sequences sampled with Gibbs sampling using the three Hamiltonians illustrated above. Each dataset was composed of 1000 sequences for every value pf distance to TEM-1 up to 30. We clearly see a trend in Fig.25, sequences that introduce more "random" mutations are penalized by our model, that is they are assigned a higher energy. Sequences generated intead by sampling the Boltzmann distribution with the Potts Hamiltonian present almost no trend in terms of increased energy versus number of mutations.

4.2 Modeling Fantini's experiment

In the last section we have obtained encouraging results. We have shown that sampling the sequence landscape locally around TEM-1 produces sequences well scored by our model. Crucial to that is the sampling from the Potts Hamiltonian learned on the dataset of natural sequences. Random sequences - or sequences whose mutations only respect the single column statistics of the MSA of natural sequences - increase instaed their energy as mutations are introduced.

We also know, as we have pointed out multiple times, that our energy scores well



(a) **(b)**

Figure 26: Density plot of the difference in energy versus the Hamming distance of the sequences belonging to the 12th generation of Fantini's experiment and to the relative artificially sampled sequences. The color indicates the number of sequences present in the spot.

the sequences coming from the two protein evolution experiments that we have analyzed. We now combine the ideas and try to simulate a protein evolution experiment in silico.

We choose the experiment of Fantini. To simulate the experiment we generated a MSA of artificially sampled sequences by Gibbs sampling from the Boltzmann distribution learned on the natural sequences. We fixed the number of sequences in the MSA to match that of the sequences in the 12th generation of the experiment. We also imposed that the distribution of mutations with respect to TEM-1 were the same. We fixed it in a way that for every sequence present in the 12th generation of Fantini's experiment, we had a sequences with the same number of mutations (compared to TEM-1) in our silico dataset.

Energy of the sequences

In analysing the sequences that we have generated, we start by looking at the distribution of changes in energy versus mutations. Confronting the two plots of Fig.25 we clearly see a good agreement. The sampled sequences occupy a slightly wider interval of energies, but overall the range of energies occupied is the same. The slope is also very similar, however the result could be due to chance, as we have reported a quite different slope Sander's sequences. Overall it seems that - as far as our fitness score is concerned - the dataset of sequences that we generated is quite similar to that of a real protein evolution experiments.

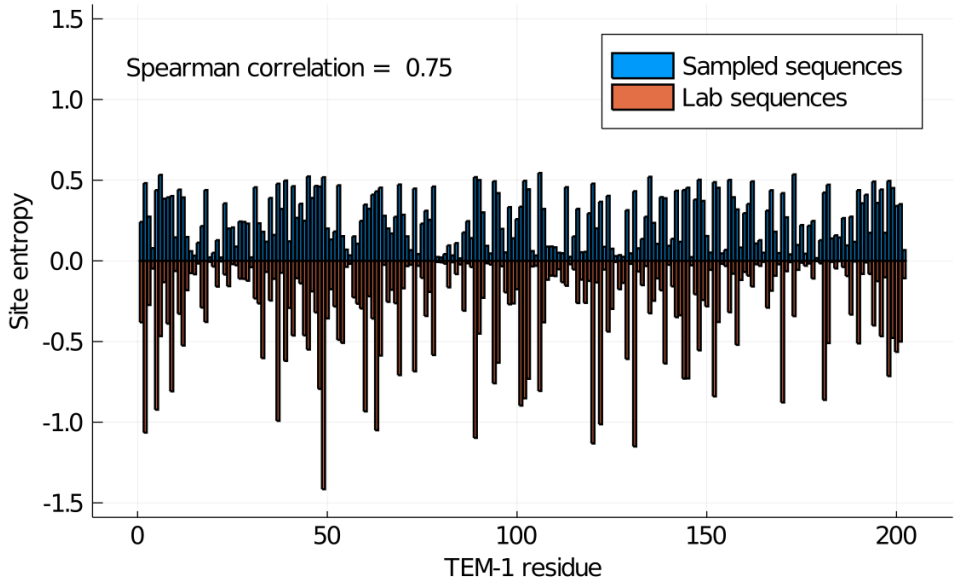


Figure 27: Shannon entropies of the distribution of the amino acids for every site of the MSAs of generation 12 of Fantini's experiment and in silico generated sequences.

Site entropy

To further our comparative analysis we look at the distribution of mutations. In particular we focus on the residue specific entropies observed in our dataset of sequences. To gauge our results we compare them to the entropies of generation 12 . We note immediately that our entropies never exceed ~ 0.5, that is we have an effective number of ~ 4.5 different amino acids that typically populate those sites. This is at odds with the entropies of Fantini's experiment, ofter much more pronounced. What is remarkable here, execially for a biology experiment, is the high value (~ 0.75) of the Spearman correlation between the entropies of the sites of our simulation and the experiment. We capture correctly the site amio acid variablity in this local landscape.

Contact prediction

We report also the contact prediction map obtained by running plmDCA on the dataset of artificial sequences. We see that we are not able to retrieve contacts between residues, as already reported in the experiment of Fantini. The result is not so unexpected, sequence divergence and variability are rather limited if compared to that of natural sequences.

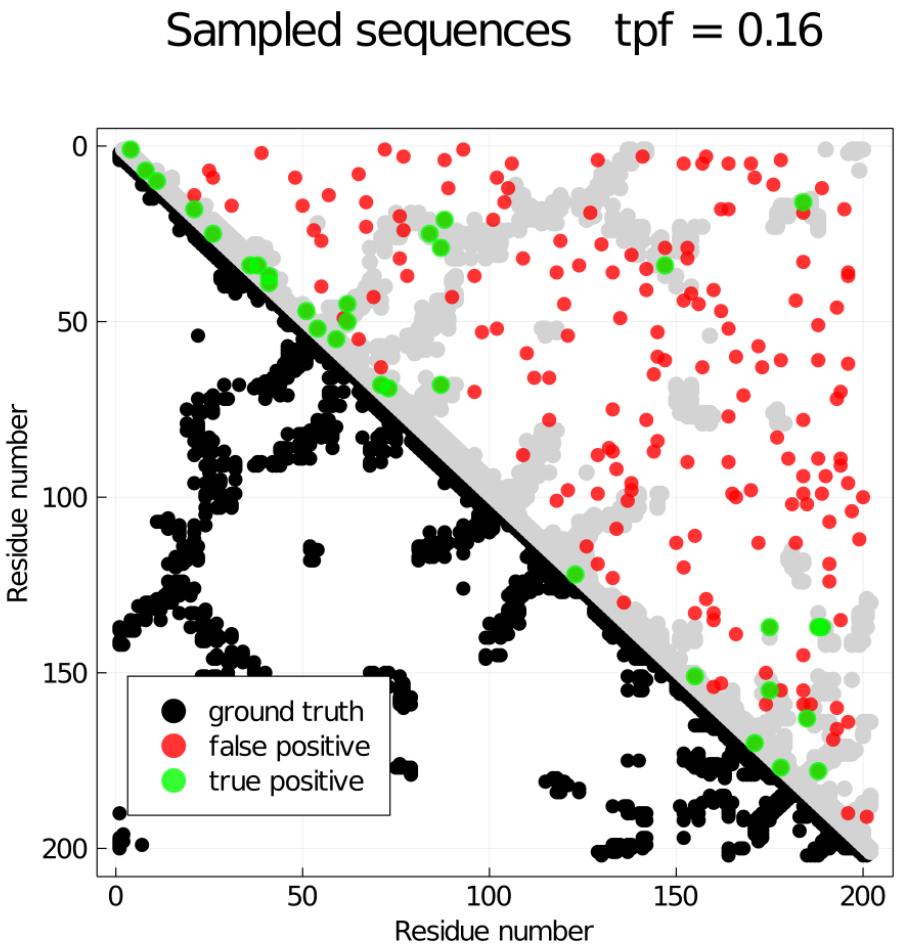


Figure 28: Contact map prediction for artificially generated sequences.

Acknowledgments

First and foremost I want to thank my supervisors, Martin and Francesco. I have learned so much from them and I can not wait to pursue my future PhD under their guide. They have been a constant stimulus and a great help expecially when I needed it the most. Their humanity, support and spontaneity have been invaluable to stay in touch and help me navigate my way though my internship while being hundreds of kilometers of distance from the lab. Among all, they have always been available.. during a pandemic.. with kids at home!

My work has benefited a lot from useful discussions with Edoardo and Juan. They have helped me find my way though the jungle of biology with patience and care. I would also like to thank all the people of Martin's group at LCQB for welcoming me warmly and naturally.

I would like also to thank Alfredo, his lessons have started in me the interest for inference and learning. He has also been very supportive during my internship and in particular during the final phases of writing.

A special thanks goes to all my class mates of Physics of Complex Systems, if I stand here now it is thanks to our collective effort, friendship and support during those two hard, but wonderful years. In particular I would like to thank Chiare, she has been a companion and friend in my best and worst moment.

My final thanks goes to my family. I owe everything to their love and support and their faith in me is invaluable.

References

[1] Marco Fantini, Simonetta Lisi, Paolo De Los Rios, Antonino Cattaneo, and Annalisa Pastore. Protein Structural Information and Evolutionary Landscape by In Vitro Evolution. *Molecular Biology and Evolution*, 37(4):1179–1192, 10 2019.

[2] Michael A. Stiffler, Frank J. Poelwijk, Kelly P. Brock, Richard R. Stein, Adam Riesselman, Joan Teyra, Sachdev S. Sidhu, Debora S. Marks, Nicholas P. Gauthier, and Chris Sander. Protein structure from experimental evolution. *Cell Systems*, 10(1):15 – 24.e5, 2020.

[3] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[4] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[5] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280, 10 2015.

[6] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. Evolution-based design of chorismate mutase enzymes. *bioRxiv*, 2020.

[7] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[8] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

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

[10] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 12 2015.

[11] William Russ, Drew Lowery, Prashant Mishra, Michael Yaffe, and Rama Ranganathan. Natural-like function in ww domains. *Nature*, 437:579–83, 10 2005.

[12] Michael Socolich, Steve Lockless, William Russ, Heather Lee, Kevin Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437:512–8, 10 2005.

[13] Robert White, Hendrik Szurmant, James Hoch, and Terence Hwa. Features of protein–protein interactions in two-component signaling deduced from genomic libraries. *Methods in enzymology*, 422:75–101, 02 2007.

[14] Jeffrey Skerker, Barrett Perchuk, Albert Siryaporn, Emma Lubin, Orr Ashenberg, Mark Goulian, and Michael Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133:1043–54, 07 2008.

[15] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[16] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 10 1998.

[17] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[18] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341–356, 2014.

[19] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[20] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios Pavlopoulos, David Kim, Hetunandan Kamisetty, Nikos Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355:294–298, 01 2017.

[21] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Phys. Rev. E*, 87:012707, Jan 2013.

[22] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 12 2007.

[23] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):1–20, 12 2011.

[24] Elizabeth Cirulli and David Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics*, 11:415–25, 06 2010.

[25] Jacob D. Mehlhoff, Frank W. Stearns, Dahlia Rohm, Buheng Wang, Erh-Yeh Tsou, Nisita Dutta, Meng-Hsuan Hsiao, Courtney E. Gonzalez, Alan F. Rubin, and Marc Ostermeier. Collateral fitness effects of mutations. *Proceedings of the National Academy of Sciences*, 117(21):11597–11607, 2020.

[26] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, 01 2018.

[27] David M Livermore and Neil Woodford. The β-lactamase threat in enterobacteriaceae, pseudomonas and acinetobacter. *Trends in microbiology*, 14(9):413–420, 2006.

[28] CDC (Centers for Disease Control and Prevention). Antibiotic resistance threats in the United States. 2019.

[29] O'Leary et. al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, Jan 2016.

[30] Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6):1581–1592, 02 2014.

[31] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

[32] Tobias Sikosek and Hue Chan. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*, 11:20140419, 08 2014.

[33] Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876, 2009.

[34] Jesse D. Bloom, Sy T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006.

[35] Bryan C Dickinson, Aaron M Leconte, Benjamin Allen, Kevin M Esvelt, and David R Liu. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proceedings of the National Academy of Sciences*, 110(22):9007–9012, 2013.

[36] Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

[37] David S Wilson and Anthony D Keefe. Random mutagenesis by pcr. *Current protocols in molecular biology*, 51(1):8–3, 2000.