POLITECNICO DI TORINO

Master's Degree in Physics of Complex Systems

Master's Degree Thesis

Maximum entropy modeling for inference in biological sequences analysis

Supervisor

Prof. Andrea PAGNANI

Candidate

Matteo DE LEONARDIS

April 2020

Acknowledgements

I would like first to thank my supervisor Prof. Andrea Pagnani who guided me through this work, contributed to fuel the interest in my studies and motivated me when difficulties arised making sure I had my doubts cleared and being always available to answer my questions.

In addition I would like to thank Dr. Guido Uguzzoni who has always been willing to help me, in particular with some coding issues and in understanding the previous work.

I would like also to acknowledge them for the consistent effort which has been been put into their work, together with Dr. Jorge Fernandez-de-Cossio-Diaz. They made available all the datasets used in this thesis and implemented all the necessary pre-processing steps carried on them, and set up the whole statistical framework on which this work is based.

In addition I would like to thank Valeria who has been with me during difficult times and accompanied me in this important journey with her loving support.

Finally I want to thank my family for the confidence and their fundamental help.

Summary

Likelihood maximization and entropy maximization are two common techniques used to infer the set of parameters of a probability distribution. In recent years, they have shown outstanding performance in inference problems of structural biology from sequence data. My work addresses two main aspects related to this subject. The first one is the prediction of contacts in a protein family through the analysis of correlation between residues. Standard information theory related methods based on local correlation measures (e.g. Mutual Information) that are routinely used to evaluate the correlation between two random variables, often fail because they are not able to disentangle direct from indirect interaction between variables. For this purpose, global inference strategies such as entropy maximization, can be used to define a quantity called "direct information" which is capable to ignore statistical correlation between residues which are not linked to the presence of contacts between them. The second research direction undertaken in my thesis, is about a maximum likelihood strategy to model phage display experiments. Phage display is a widespread laboratory technique (2018 Nobel prize in Chemistry) for the study of protein–protein, protein–peptide, and protein–DNA interactions that uses bacteriophages (viruses that infect bacteria) to connect proteins with the genetic information that encodes them. A coding gene is inserted into the phage genome to expose the protein under study on the phage capsid. Typically, a population of $10^{1}3$ phages is grown to display variants of wild-type proteins encoded in biologically engineered combinatorial libraries. This allows for screening tests, repeated for a certain number of rounds, aimed at testing their binding capability against a target. After each round, the phage population can be sequenced to inspect the abundance of sequences that are bound to the target. Usually, supervised machine learning approaches are utilized to analyze phage display experiments in order to predict the selectivity of new sequences. Nevertheless, an unsupervised approach based on Likelihood Maximization

can be developed by outlining a model based on statistical mechanics which describes the experiment and it allows for the statistical inference of the relevant parameters of the model. This is carried out through a multi-variate optimization of a likelihood score. Thanks to this approach, the binding of the sequence to the target is modelled in a probabilistic way in terms of a two-states system by using an "energy" function that depends on the amino acid sequence. Finally, this model can be extended to a three-states system in which the third state can be associated to the state in which the sequence is folded but still cannot bind to the target.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction: The basic tools

1.1 Fitness Landscapes: a general framework

Due to technical advances in experimental techniques larger and larger libraries of sequences can be sequenced and this has caused an outburst of experimental data. This abundance calls for novel and more efficient methods to analyze this large amounts of data. Statistical physics has provided the fundamental tools to provide this methods and it has revealed as a perfect match with these experimental techniques to devise general approach aimed at giving insights into biological sequences properties and pave the way to solve difficult and demanding problems in structural biology.

The key strategy to design these combined approaches is to study the properties of sequences by creating large libraries of mutated variants of a specific sequence and to observe how these mutations affect the property under study (e.g. the 3D structure in sec. 2 or its binding affinity in sec. 3). This is something that, in a certain way, mimics natural selection and thanks to these powerful experimental techniques one can reconstruct the general landscape that describes how the specific properties changes by changing the sequences.

Statistical physics has proven to be a very powerful tool to design efficient models to derive this fitness landscape, this because in general, at some point, some stochasticity enters in in order to provide an accurate description of the processes involved in the experimental counterpart.

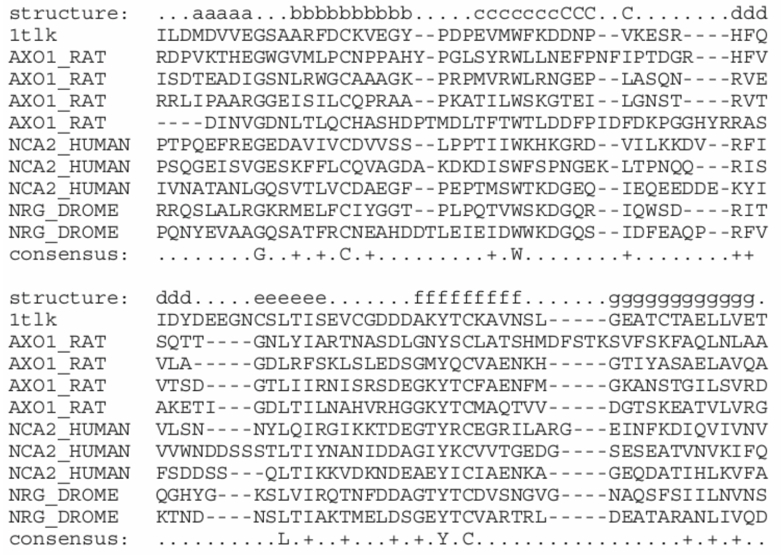
1.2 A probabilistic model for the data

[1] When analyzing sequences the property under study can be explored by grouping sequences in sets of homologous: groups of sequences that can be considered as related, for instance because one can think that they have mutated from a common ancestor, typically because they are quite similar. That being said, a very general approach to this problem is to construct a set of homologous sequences that is in general composed by a wild type and a certain number of sequences that are similar to that particular one because these sequences have mutated from that wild type (these sequences are usually called *variants*) and the property under study can be tested on these different sequences to outline the precise part of the sequence that is responsible for it.

Detecting differences between two sequences can be thought as the same as finding differences between two text strings because one can consider the primary structure of a sequence as a sequence of symbols (e.g. aminoacids in proteins, nitrogenous bases in DNA strands). This problem of assessing the similarity between two sequences is a well known problem in computer science and it is tackled by modeling mutations as a composition of three possible operations that can be done on strings: the insertion of a character, the deletion of a character and the substitution of a character with another; then one can count how many of these operation must be done on a sequence to obtain another one and score their amount of similarity.

Based on this approach one can quantify the similarity between two sequences and attribute a score to them. This scoring system can be used not only to assess whether two sequences can be considered similar, but also to reconstruct the most likely sequence of elementary operation that produced that particular sequence by means of evolution, and this can be represented by a multiple sequences alignment (MSA), an example is shown in figure 1.1. The task of building these sets of homologous and to align them is in general a difficult one, and even if much progress has been already carried out on this topic, it is still an open one that should be addressed by itself. For this reason in the proceeding of this work it will be assumed that a MSA of a family is given.

In generalizing this approach, when one wants to assess a property of a certain sequence, different from the occurrence of a residue in a certain position, a more sophisticated probabilistic model must be employed in order to rely on a meaningful inference after observing some data and avoiding



```
structure:  ...aaaaa...bbbbbbbbbb.....cccccccCCC..C........ddd
1tlk        ILDMDVVEGSAARFDCKVEGY--PDPEVMWFKDDNP--VKESR----HFQ
AXO1_RAT    RDPVKTHEGWGVMLPCNPPAHY-PGLSYRWLLNEFPNFIPTDGR---HFV
AXO1_RAT    ISDTEADIGSNLRWGCAAAGK--PRPMVRWLRNGEP--LASQN----RVE
AXO1_RAT    RRLIPAARGGEISILCQPRAA--PKATILWSKGTEI--LGNST----RVT
AXO1_RAT    ----DINVGDNLTLQCHASHDPTMDLTFTWTLDDFPIDFDKPGGHYRRAS
NCA2_HUMAN  PTPQEFREGEDAVIVCDVVSS--LPPTIIWKHKGRD--VILKKDV--RFI
NCA2_HUMAN  PSQGEISVGESKFFLCQVAGDA-KDKDISWFSPNGEK-LTPNQQ---RIS
NCA2_HUMAN  IVNATANLGQSVTLVCDAEGF--PEPTMSWTKDGEQ--IEQEEDDE-KYI
NRG_DROME   RRQSLALRGKRMELFCIYGGT--PLPQTVWSKDGQR--IQWSD----RIT
NRG_DROME   PQNYEVAAGQSATFRCNEAHDDTLEIEIDWWKDGQS--IDFEAQP--RFV
consensus:  ........G..+.+.C.+.........+.W........+........++

structure:  ddd.....eeeeee......fffffffff.......gggggggggggg.
1tlk        IDYDEEGNCSLTISEVCGDDDAKYTCKAVNSL-----GEATCTAELLVET
AXO1_RAT    SQTT----GNLYIARTNASDLGNYSCLATSHMDFSTKSVFSKFAQLNLAA
AXO1_RAT    VLA-----GDLRFSKLSLEDSGMYQCVAENKH-----GTIYASAELAVQA
AXO1_RAT    VTSD----GTLIIRNISRSDEGKYTCFAENFM-----GKANSTGILSVRD
AXO1_RAT    AKETI---GDLTILNAHVRHGGKYTCMAQTVV-----DGTSKEATVLVRG
NCA2_HUMAN  VLSN----NYLQIRGIKKTDEGTYRCEGRILARG---EINFKDIQVIVNV
NCA2_HUMAN  VVWNDDSSSTLTIYNANIDDAGIYKCVVTGEDG----SESEATVNVKIFQ
NCA2_HUMAN  FSDDSS---QLTIKKVDKNDEAEYICIAENKA-----GEQDATIHLKVFA
NRG_DROME   QGHYG---KSLVIRQTNFDDAGTYTCDVSNGVG----NAQSFSIILNVNS
NRG_DROME   KTND----NSLTIAKTMELDSGEYTCVARTRL-----DEATARANLIVQD
consensus:  ..........L.+..+...+.+.Y.C.................+.+.+..
```

Figure 1.1: Example of MSA of ten I-set immunoglobulin domains, figure reprinted from [1].

the act of just counting how many sequences displaying a certain feature is present in the dataset.
This probabilistic model can be considered as a *stochastic machine* that produces randomly some output, and these are the samples that we get to observe. In the case of MSA the stochastic machine that is usually employed is an Hidden Markov Model that generates aligned sequences belonging to the the set of homologous sequences that we want to outline (this is a simple generalization that corresponds to adding a symbol, the gap "−" among the ones that can compose the sequences).
Introducing a probabilistic model is a tool that can employed not only to devise an efficient aligning method but also more generally to model many possible processes that the particular sequences under study can undergo, to investigate the property that can be associated to these sequences.
Of course a model can be designed to describe also different kinds of information, therefore the approach that will be exposed and analyzed in this work is indeed a very general one and it must be carefully adapted to the data that must be analyzed. For instance, in chapter 2, a technique known as Direct Coupling Analysis (DCA) will be exposed and it will be shown how it can give insight into coevolutionary patterns of homologous sequences. In chapter 3, instead, taking inspiration from DCA a similar approach is used

to analyze data coming from Phage Display experiments that are composed of sets of reads of the various sequences at each round of the experiment; this kind of data are fundamentally different from a MSA.

1.3 Maximum Entropy

Once a model that can describe the occurrence of the data has been designed, one typically wants to determine it after having observed the data. This a very old problem in statistics and many solutions has been proposed during the time. As shown in [2] it is possible to tackle this problem considering a quantity known as Shannon's entropy. Let us consider for the moment a more general framework with respect to the one that one uses when performing inference starting from a particular dataset. Consider a random variable $x \in X$ with $|X| = n$ finite and its possible outcomes are labelled as $(x_1, \cdots, x_n)$, moreover $p : X \longrightarrow R_{\geq 0}$ a probability distribution over X. One can consider the problem of determining the p_is considering as known the expectation of some functions $f_r(x)$, namely

$$\langle f_r(x) \rangle = \sum_i p_i f_r(x_i) \tag{1.1}$$

typically the knowledge of these expectation values encodes the piece of information coming from the observed data.
Shannon's entropy is defined as

$$S(p) = \sum_i p_i \log p_i \tag{1.2}$$

and it can be shown that this quantity can be associated with the *amount of uncertainty* about the outcome of the random variable x. Even if the concept of uncertainty lacks of a mathematical definition it can be shown that such a quantity comes from very simple requested properties and it encodes the fact that we consider as maximally uncertain the outcome of x when $p_i = 1/n \; \forall i$ and it is minimally uncertain when the probability is concentrated on a particular outcome, according to our intuitive meaning of uncertainty. The simple properties from which the expression of S will follow are 3:

1. $S(p)$ is a continuous functions of the p_is
2. $S(1/n, \cdots, 1/n)$ is an increasing function of n

3. Consider $\{\pi_j\}_{j=1}^k$ a partition of the set $\{1, \cdots, n\}$ such that if $l, m \in \pi_i$ and $l \leq k \leq m$ then $k \in \pi_i$, in addition $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^k \pi_j = \{1, \cdots, n\}$.
   Define $w(\pi_i)$ as the probability of the subset of outcomes π_i and naturally $w(\pi_i) = \sum_{j \in \pi_i} p_j$.
   Then define $w(x_m|\pi_i)$ as the probability of event $m \in \pi_i$ given that subset π_i is chosen and in this case

$$w(x_m|\pi_i) = \frac{p_m}{\sum_{j \in \pi_i} p_j} \tag{1.3}$$

   Now one can require that the amount of information doesn't change if the realization of event x_m happens in two steps instead of one: first a π_i is specified, then event x_m is chosen from it. In mathematical terms

$$S(\{p_j\}_{j=1}^n) = S(\{w(\pi_j)\}_{j=1}^k) + \sum_{j=1}^k w(\pi_j) S(\{w(x_m|\pi_j)\}_{m \in \pi_j}) \tag{1.4}$$

Exploiting property 1, one determine the function S only for rational values of the p_is and then the function can be extended to values in $\mathbb{R} \setminus \mathbb{Q}$. That being said one can consider p_i in the form of

$$p_i = \frac{n_i}{\sum_i n_i} \quad \text{with } n_i \in N \tag{1.5}$$

One can now notice that the p_is has assumed the form of the probability of one of the n_i events among $\sum_i n_i$ equally probable events $\{a_r\}_{r=1}^{\sum_i n_i}$. Therefore one can regard the outcome of x as a first step before one of the corresponding n_i events is chosen and setting $x_i \doteq \pi_i = \{a_{r+1}, \cdots, a_{r+n_i}\}$. Now let us define $A(n) = S(1/n, \cdots, 1/n)$ as the entropy of the uniform distribution among n possible outcomes and then from property 1.4 one can write

$$S(p_1, \cdots, p_n) + \sum_i^n p_i A(n_i) = A(\sum_{i=1}^n n_i) \tag{1.6}$$

Setting $n_i = m$ for all is and using eq. 1.5, eq. 1.6 reduces to

$$A(n) + A(m) = A(nm) \tag{1.7}$$

which is clearly solved by $A(n) = K \log n$ with $K > 0$. Now inserting this solution into eq. 1.6 the well known expression of Shannon's entropy in eq. 1.2 is obtained. Going back to consider the initial inference problem now it seems natural that the probability distribution should be the one that maximizes the entropy, the one that makes the least assumption about the system under study and therefore introduces the least amount of bias, constrained to what is considered as known from the observed data, encoded in the form of eq. 1.1.

1.4 Maximum Likelihood estimation

The inference problem considered in section 1.3 is indeed very general but still quite useless if we apply that approach without further work. This because when using machine learning to gain information about the properties of sequences we are not interested in knowing how they depend on the particular sequences in the dataset, rather on the particular primary structure of the sequence. In this way one could also use the information to infer the particular property of interest on sequences that were not observed experimentally, or to generate sequences having that particular property.

This is what is typically done when using machine learning: a large set of samples, which is supposed to be sufficiently representative of the whole space of possible sequences that can be found, is taken and it is used to learn how this specific property depends on the primary structure of the sequence and then this information is used to infer this property on any other sequence that one could be interested in studying.

This task is accomplished by writing the probability distribution of the stochastic machine that in our model generates the samples in terms of a set of parameters $\boldsymbol{\theta}$ that is used to encode the information of the primary structure of the sequence, namely $p(x|\boldsymbol{\theta})$.

A problem which is somehow complementary to the one of inferring the form of the probability distribution investigated in section 1.3 is the one of determining the best set of parameters $\boldsymbol{\theta}$ according to the observed data considering the form of the probability distribution as known.

At first sight this problem seems way easier with respect to the one of determine the form of the probability distribution, and it is often the case as will be illustrated in chapter 2 for the case of DCA, in that case this problem can be solved together with the one of determining the form of the probability distribution within the Maximum entropy approach. In other cases instead the system under study undergoes some physical processes before we can observe something about it, and neglecting this fact, when designing the probabilistic model, can lead to the incapability of the model to describe these processes and this could result in a poor performance when using the learnt information to infer something about sequences that are not observed.

When this is the case more effort must be put into working out a model that is capable to capture all relevant aspects concerning the observation of the experimental data. For example, in chapter 3, there will be the need to model the physical experiment of phage display in a probabilistic way so that the learnt information can be meaningful and can be used on any compatible data that undergoes the same physical processes before being observed.

Since this is an approach that relies deeply on the particular problem that one wants to study there is no general recipe to design such a model but the problem must be tackled in a specific way depending on the particular system under study. The only general principle that can be employed in this task is the one of Maximum Likelihood estimation [3], or more specifically of Maximum Aposteriori estimation, that comes very naturally from a Bayesian point of view of the problem that can be expressed with the following simple expression resulting from a simple application of Bayes's Theorem:

$$p(\boldsymbol{\theta}|\text{data}) \propto p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.8}$$

were $p(\boldsymbol{\theta})$ is any probability distribution over parameters $\boldsymbol{\theta}$ coming from any prior knowledge about these parameters and the likelihood $p(\text{data}|\boldsymbol{\theta})$ is the probability of our model to generate the observed data. And it is now natural to look for the $\boldsymbol{\theta}$ that maximizes $p(\boldsymbol{\theta}|\text{data})$. Sometimes this approach passes through additional variables and parameters that can be used to model the physical processes of the experiment, and this will be the case of phage display and its modeling process will be addressed in chapter 3.

Chapter 2

Direct Information

2.1 Mutations at contact points

Proteins assume a well defined structure in order to perform their biological tasks and a particular mutation at a certain point of its chain can in principle alter its functionality. For this reason, nature selection accepts only mutations that preserve the protein structure and allows them to continue to carry out their particular tasks. Anyway it is experimentally observed that two proteins that have a common ancestor have the same three-dimensional structure but they differ from 70-80% in their amino acids [4].

One can imagine that these processes can happen in two different steps: first an harmful single site mutation occurs and than a second mutation happens at the corresponding site at which contact has been lost, and this second mutation occurs in such a way to restore the correct contact point necessary to maintain the correct 3D structure.

Assuming that a MSA of proteins of a certain family is already been given, one could expect to find correlation between residues that form contact points. Unfortunately this approach to reconstruct contact points doesn't work since sites along the chain that are highly correlated cannot be associate directly to contact points. The reason is that correlation between two sites may arise because of small correlation contributes between intermediary sites without any presence of contact [5].

2.2 Local measure

2.2.1 Frequency counts

In order to detect correlation it is necessary to count one-site and double-site residue occurrences. In particular being a_i^m the residue of the $m-th$ sequence occurring at site i with $m = 1, \cdots, M$ and $i = 1, \cdots, L$, one can compute occurrences frequencies:

$$f_i(a) = \frac{1}{M} \sum_m \delta_{a,a_i^m} \tag{2.1}$$

$$f_{ij}(a,b) = \frac{1}{M} \sum_m \delta_{a,a_i^m} \delta_{b,a_j^m} \tag{2.2}$$

2.2.2 Re-weighting sequences

Before starting with the analysis of correlation between the various couples of sites, a first step is needed to deal with the non uniformity of sequence sampling inside a family. Since in a protein family we expect many similarities due to the fact that they come from a common ancestor, we need this first step because, otherwise, one could detect correlation between sites which are not linked to a coupling.

Another reason why one should implement this first re-weighting step is that often, sequences are sampled due to their relevance in various studies and sequences that are not involved in important roles are less frequently sequenced in databases and this can introduce bias in counting as well. Let us fix a threshold $x \in (0{,}1)$ and use the Hamming distance (portion of distinct residues between two sequences) to define if two sequences can be considered identical, in practice they bring the same information to our analysis.

Let us define the number of sequences "identical" to a certain one

$$n^m = |\{b|1 \leq b \leq M, \ \text{seqid}(A^m, A^b) \geq xL\}| \tag{2.3}$$

where A^a represents the a-th sequence of MSA and seqid is the funciton that measures the percentage of match between two sequences.

Let us now use this quantity to define the counting weight $w_m = 1/n^m$ of the sequence a. Now this weight can be inserted in 2.1-2.2 turning them into

$$f_i(a) = \frac{1}{M_{eff}} \sum_m w_m \delta_{a,a_i^m} \tag{2.4}$$

$$f_{ij}(a,b) = \frac{1}{M_{eff}} \sum_m w_m \delta_{a,a_i^m} \delta_{b,a_j^m} \tag{2.5}$$

where

$$M_{eff} = \sum_m w_m \tag{2.6}$$

Let us investigate the effect of this re-weighting step.
Eq. 2.3 divides sequences in k classes of identical sequences

$$C_j = \{A^m | 1 \leq m, m' \leq M,\ \text{seqid}(A^m, A^{m'}) \leq xL\ \forall m' \text{ s.t. } A^{m'} \in C_j\} \tag{2.7}$$

with $j = 1, \cdots, k$, and let us consider ,for a generic residue a, the term

$$\sum_m w_m \delta_{a,a_i^m} = \sum_{j=1}^{K} \sum_{m \in C_j} \frac{1}{n^m} \delta_{a,a_i^m} = \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{m \in C_j} \delta_{a,a_i^m} \tag{2.8}$$

The last term $\frac{1}{|C_j|} \sum_{m \in C_j} \delta_{a,a_i^m}$ in the previous equation is the average frequency count for a in position i in the j-th class. And moreover

$$M_{eff} = \sum_m w_m = \sum_{j=1}^{K} \sum_{m \in C_j} \frac{1}{n^m} = \sum_{j=1}^{K} 1 = K \tag{2.9}$$

Now one can easily see that $f_i(a)$ becomes

$$f_i(a) = \frac{1}{K} \sum_{j=1}^{K} \text{average frequency count of } a \text{ in position } i \text{ in class } j \tag{2.10}$$

This re-weighting steps gives equal weights to all classes of identical sequences in MSA in the computation of empirical counts. Exactly the same con be said about two-sites frequency counts getting to 2.5.
Setting x to 1 would simply re-weight sequences that are sampled more than once, and lower values of x are used to correct more imbalanced MSA.

2.2.3 Pseudo-counts

In order to correct finite-size effect of the MSA one could introduce pseudo-counts. Let us say that a total number of λ pseudo-counts are introduced for all residues, it means that the counts for each residue start from λ/q and, whatever type of re-weighting is chosen, Eq. 2.1 becomes

$$f_i(a) = \frac{1}{\lambda + M}\left(\frac{\lambda}{q} + \sum_m \delta_{a,a_i^m}\right) \tag{2.11}$$

When introducing pseudo-counts in two-sites frequencies, instead, a bit of attention must be paid because it is necessary to keep consistency between one-site and two-sites frequencies, because by marginalizing one obtains

$$\sum_b f_{ij}(a,b) = f_i(a) \tag{2.12}$$

Therefore the correct way to introduce pseudo-counts is

$$f_{ij}(a,b) = \frac{1}{\lambda + M}\left(\frac{\lambda}{q^2} + \sum_m \delta_{a,a_i^m}\delta_{b,a_j^m}\right) \tag{2.13}$$

In fact by marginalizing one gets

$$\begin{aligned}\sum_b f_{ij}(a,b) = \sum_b \frac{1}{\lambda + M}\left(\frac{\lambda}{q^2} + \sum_m \delta_{a,a_i^m}\delta_{b,a_j^m}\right) = \\ \frac{1}{\lambda + M}\left(\frac{\lambda}{q} + \sum_m \delta_{a,a_i^m}\right) = f_i(a)\end{aligned} \tag{2.14}$$

In this approach the occurrence of a certain residue in a position is modelled through a categorical probability distribution, and the probability of observing residue a at a site i is given by $p(a) = \theta_i(a)$ where $\theta_i(a)$ is the real probability of occurrence of a at site i and $\sum_b \theta_i(b) = 1$ must hold for every $i = 1, \cdots, L$. In this framework the probability of observing a certain sequence is given by

$$p(A^m|\boldsymbol{\theta}) = \prod_{i=1}^{L} \theta_i(a_i^m) \tag{2.15}$$

and the likelihood of the MSA can e written as

$$p(A|\boldsymbol{\theta}) = \prod_{i=1}^{L}\prod_{a=1}^{q} \theta_i(a)^{N_i(a)} \tag{2.16}$$

where $N_i(a) = \sum_m \delta_{a,a_i^a}$ is the number of times the residue a appears at site i in the MSA. In a Bayesian framework in which

$$p(\boldsymbol{\theta}|A) \propto p(A|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{2.17}$$

one can interpret pseudo-counts as a Dirichelet prior over the space of admissible parameters $\boldsymbol{\theta}$: the simplex given by $\prod_i S_i$ where $S_i = \{\boldsymbol{\theta} : \sum_b \theta_i(b) = 1\}$, and $p(\boldsymbol{\theta})$ is given by

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{L} \prod_{a=1}^{q} \theta_i(a)^{\lambda/q} \tag{2.18}$$

2.2.4 Mutual Information

Once one-site and two-sites frequencies have been computed, Mutual Information (MI), a standard method to compute correlation in information theory, can be used to detect correlation in occurrences at various pairs of sites, namely:

$$M_{ij} = \sum_{a,b} f_{ij}(a,b) \log\left(\frac{f_{ij}(a,b)}{f_i(a)f_j(b)}\right) \tag{2.19}$$

the KL divergence between f_{ij} and f_i^2 (point-wise multiplication). M_{ij} vanishes when $f_{ij}(a,b) = f_i(a)f_j(b)$ for each couple (a,b).
The problem whith this method is that while it is expected that a contact between sites i and j results in a not small value of M_{ij} the converse is not true in general because, as already said in 2.1 correlation between sites may arise due to indirect coupling between intermediate sites.

2.3 Global probability distribution

In order to disentangle direct from indirect correlation one can adopt a strategy which is different from the one described in the previous section. Instead of computing the mutual information of a local distribution (one/two-site estimates) one could infer a global distribution $P(a_1, \cdots, a_L)$ and obtain single and double-site frequencies as marginals of this global distribution

$$P_i(a_i) = \sum_{\boldsymbol{a}\setminus\{a_i\}} P(a_1, \cdots, a_L) \tag{2.20}$$

$$P_{ij}(a_i, a_j) = \sum_{\boldsymbol{a}\setminus\{a_i,a_j\}} P(a_1, \cdots, a_L) \tag{2.21}$$

This global distribution con be inferred in two ways but they are equivalent. The first, more general method, is to use entropy maximization; the second one assumes that sequences are drawn from a Boltzmann probability distribution with an Hamiltonian of an Ising model with pair couplings.

2.3.1 Maximum entropy modelling

From Information theory it is known that entropy can be associated to the "quantity of information" known for a certain system and it suggests that the probability distribution with maximum entropy is the most unbiased one (which retains the biggest amount of information) that one can use.
Besides that the entropy must be maximized, one should request that marginals in Eqs. 2.20 and 2.21 match empirical one-site and two-sites frequencies for each i, j, a_i, a_j.
Adding another constraint due to the normalization of the distribution, the Lagrangian will be given by

$$\begin{aligned}\mathcal{L}[P] = &-\sum_{\boldsymbol{a}} P(a_1,\cdots,a_L)\log P(a_1,\cdots,a_L)+\sum_i\sum_{a_i} h_i(a_i)(P_i(a_i)-f_i(a_i))\\ &+\sum_{i<j}\sum_{a_i,a_j} e_{ij}(a_i,a_j)(P_{ij}(a_i,a_j)-f_{ij}(a_i,a_j))+Q(\sum_{\boldsymbol{a}} P(a_1,\cdots,a_L)-1) =\\ &\sum_{\boldsymbol{a}}\Big\{-P(a_1,\cdots,a_L)\log P(a_1,\cdots,a_L)+\sum_i h_i(a_i)(P(a_1,\cdots,a_L)-f_i(a_i))\\ &+\sum_{i<j} e_{ij}(a_i,a_j)(P(a_1,\cdots,a_L)-f_{ij}(a_i,a_j))+Q(P(a_1,\cdots,a_L)-\frac{1}{q^L})\Big\}\end{aligned} \tag{2.22}$$

The $P(a_1,\cdots,a_L)$ that maximizes $\mathcal{L}[P]$ must satisy the stationarity condition given by

$$\begin{aligned}&\frac{\delta}{\delta P(a_1,\cdots,a_L)}\mathcal{L}[P] =\\ &\qquad -\log P(a_1,\cdots,a_L)+\sum_i h_i(a_i)+\sum_{i<j} e_{ij}(a_i,a_j)+const=0\end{aligned} \tag{2.23}$$

and therefore one can obtain

$$P(a_1,\cdots,a_L)=\frac{1}{Z}\exp\left\{\sum_{i<j} e_{ij}(a_i,a_j)+\sum_i h_i(a_i)\right\} \tag{2.24}$$

where Z is given by normalization

$$Z = \sum_{\boldsymbol{a}} \exp\left\{\sum_{i<j} e_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right\} \tag{2.25}$$

Even though the form of the distribution $P(a_1, \cdots, a_L)$ can be obtained from the necessary condition of stationarity the actual values of the Lagrange multipliers $\{e_{ij}(a_i, a_j)\}$ and $\{h_i(a_i)\}$, and Z (in place of the Lagrange multiplier Q which, basically, plays the same role) must still be computed and there are many algorithms that can be used to address this problem.

2.3.2 Gauge fixing

When inferring the probability distribution in Eq.2.24 in principle one should determine $Lq + \frac{1}{2}L(L-1)q^2$ different parameters but a careful analysis shows that not all of the are independent. Basically because constraint on single and double site frequencies must be consistent through Eq. 2.14 and their normalization. In detail it must be that

$$\sum_a f_i(a) = 1 \quad \forall i \tag{2.26}$$

and therefore only $L(q-1)$ of them are independent. For double-site frequencies

$$\sum_b f_{ij}(a, b) = f_i(a) \tag{2.27}$$

must hold for every i; and therefore only $q-1$ of them are independent. But it must also hold that

$$\sum_a f_i j(a, b) = f_j(b) \tag{2.28}$$

for every j and again only $q-1$ of them are independent. This means that the number of independent couplings is $\frac{1}{2}L(L-1)(q-1)^2$. The remaining parameters can in principle be fixed arbitrarily (this procedure in physics is known as Gauge fixing and this is the reason why the same name is used) as if all energy values are measure with respect to a certain residue chosen as reference .

For instance a possible Gauge fixing is the following in which the gap is chosen as reference

$$e_{ij}(a, -) = e_{ij}(-, a) = h_i(-) = 0 \quad \forall a, \forall i < j \tag{2.29}$$

2.4 Direct Information

At this point, a quantity called Direct Information (DI) can be introduced in order to detect correlation between sites filtering out correlation due to intermediate residues which cannot be linked structural contacts.
Considering a two-sites model which is composed only of sites i and j, let us define a probability distribution on this model given by

$$P_{ij}^{(dir)}(a,b) = \frac{1}{Z_{ij}} \exp\left(e_{ij}(a,b) + \tilde{h}_i(a) + \tilde{h}_j(b)\right) \tag{2.30}$$

where Z_{ij} is a normalization factor and the two fields $\tilde{h}_i(a)$ and $\tilde{h}_j(b)$ must be chosen in order to satisfy consistency with single-site frequencies, namely

$$\sum_b P_{ij}^{(dir)}(a,b) = f_i(a) \tag{2.31}$$

$$\sum_a P_{ij}^{(dir)}(a,b) = f_j(b) \tag{2.32}$$

At this point DI can be introduced as the mutual information of this two-sites distribution respect to single-site frequencies

$$DI = \sum_{a,b} P_{ij}^{(dir)}(a,b) \log\left(\frac{P_{ij}^{(dir)}(a,b)}{f_i(a)f_j(b)}\right) \tag{2.33}$$

This quantity is expected to be smaller than MI because the coupling between i and j is the only one present in this new model and it vanishes when there are no couplings, $e_{ij}(a,b) = 0$ for each a, b, because of constraints in Eq. 2.31.

2.5 Maximum Likelihood

Another way to infer the global distribution $P(a_1, \cdots, a_L)$ is to use a maximum likelihood approach [4].
Assuming that sequences are drawn from a Boltzmann distribution with an Hamiltonian given by an Ising model with pair couplings, one can find the parameters $\{e_{ij}(a,b)\}$ and $\{h_i(a)\}$ by finding the ones that maximise the (weighted, if a re-weighting if the sequences has been performed as described in section 2.2.2) (log-)likelihood

$$\mathcal{L} = \frac{1}{M_{eff}} \sum_m w_m \log P(a_1^m, \cdots, a_L^m) \tag{2.34}$$

where

$$P(a_1, \cdots, a_L) = \frac{1}{Z} \exp\left(\sum_{i<j} e_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right) \tag{2.35}$$

Re-expressing Eq. 2.34 one obtains

$$\begin{aligned}\mathcal{L} = \frac{1}{M_{eff}} \sum_m w_m \log P(a_1^m, \cdots, a_L^m) = \\ \frac{1}{M_{eff}} \sum_m w_m [-\log Z + \sum_{i<j} e_{ij}(a_i^m, a_j^m) + \sum_i h_i(a_i^m)] = \\ -\log Z + \frac{1}{M_{eff}} \sum_m w_m \sum_{i<j} \sum_{a,b} \delta_{a_i^m, a} \delta_{a_j^m, b} e_{ij}(a, b) + \frac{1}{M_{eff}} \sum_m w_m \sum_i \sum_a \delta_{a_i^m, a} h_i(a) \\ = -\log Z + \sum_{i<j} \sum_{a,b} e_{ij}(a, b) f_{ij}(a, b) + \sum_i \sum_a h_i(a) f_i(a) \end{aligned} \tag{2.36}$$

If we subtract to this quantity the entropy of the empirical distribution: the one that has, in principle, the single-site and double-site frequency counts as marginals, $f(\boldsymbol{b}) = \frac{1}{M_{eff}} \sum_m \prod_{i=1}^{L} \delta_{b_i, a_i^m}$ such that

$$f_i(a_i) = \sum_{\boldsymbol{a} \setminus \{a_i\}} f(\boldsymbol{a}) \quad \forall i \tag{2.37}$$

$$\text{and} \tag{2.38}$$

$$f_{ij}(a_i, a_j) = \sum_{\boldsymbol{a} \setminus \{a_i, a_j\}} f(\boldsymbol{a}) \quad \forall i < j \tag{2.39}$$

one obtains the the opposite of the KL divergence of $f(\boldsymbol{a})$ with respect to $P(\boldsymbol{a})$. Therefore the optimal values of the parameters can be seen as the ones that minimizes the KL-divergence

$$D_{KL}(f||P) = \sum_{\boldsymbol{a}} f(\boldsymbol{a}) \log \frac{f(\boldsymbol{a})}{P(\boldsymbol{a})} \tag{2.40}$$

In fact

$$D_{KL}(f||P) = \sum_{\boldsymbol{a}} \left(\underbrace{f(\boldsymbol{a}) \log f(\boldsymbol{a})}_{(*)} \underbrace{- f(\boldsymbol{a}) \log P(\boldsymbol{a})}_{(**)} \right) \tag{2.41}$$

the term labeled as $(*)$ is the entropy of $f(\boldsymbol{a})$ and the term labelled as $(**)$ can be expressed as

$$-\sum_{\boldsymbol{a}} f(\boldsymbol{a}) \log P(\boldsymbol{a}) = -\sum_{\boldsymbol{a}} f(\boldsymbol{a}) \left(-\log Z + \sum_{i<j} e_{ij}(a_i, a_j) + \sum_{i} h_i(a_i) \right) = \\ = \log Z - \sum_{i<j} \sum_{a_i, b_i} e_{ij}(a_i, b_i) f_{ij}(a_i, a_j) - \sum_{i} \sum_{a_i} h_i(a_i) f_i(a_i) \quad (2.42)$$

which is exactly the quantity in Eq. 2.36 with opposite sign.

2.5.1 Equivalence of the two methods

It can be easily seen that the two approaches are equivalent and bring to the same optimal set of parameters $\{e_{ij}(a,b)\}$ and $h_i(a)$.
In fact, stationarity of the constrained entropy implies the form of the distribution given by Eq. 2.24, and inserting it back in the constrained entropy to be maximised one obtains

$$S = -\sum_{\boldsymbol{a}} P(a_1, \cdots, a_L) \log P(a_1, \cdots, a_L) + \sum_{i} \sum_{a_i} h_i(a_i)(P_i(a_i) - f_i(a_i)) \\ + \sum_{i<j} \sum_{a_i, a_j} e_{ij}(a_i, a_j)(P_{ij}(a_i, a_j) - f_{ij}(a_i, a_j))) = \\ \log Z + \sum_{ij} \sum_{a_i, a_j} e_{ij}(a_i, a_j) f_{ij}(a_i, a_j) + \sum_{i} \sum_{a_i} f_i(a_i) \quad (2.43)$$

which apart from the different term $\log Z$, which is a constant independent of the couplings and fields, is the same expression of Eq. 2.36 and it is maximised by the same set of parameters $\{e_{ij}(a,b)\}$ and $\{h_i(a)\}$. This time, in the expression of the constrained entropy, the normalization has not been included because it is already taken into account in the expression of the Boltzmann distribution of Eq. 2.35.

Chapter 3

Phage Display

Taking inspiration from the method exposed in Chapter 2 it is possible to develop a statistical model which can be used to analyze data coming from experiments aimed at testing the affinity of certain peptides to a given target and predict this capability for peptides that where never observed. One relevant technique is Phage display that was first tested by George P. Smith in 1985 and the developed and enhanced by Greg Winter (and John McCafferty) which got them the 2018 Nobel Prize in Chemistry.

3.1 Brief overview of the experiment

This technique consists in taking some genes coding for a certain peptide and inserting them in the DNA strand of a particular type of virus called *bacteriophage*. This particular virus is capable to expose the corresponding peptide on its outside just like a “hat”, in this way one phage carries the information about the connection between genotype and phenotype, and being the peptide exposed on its hat it is possible to test its affinity towards a certain target molecule. Using a protein-engineering technique called *directed evolution*, developed by Frances Arnold (2018 Nobel Prize in Chemistry), it is possible to create a large library of phages carrying mutated genes of a wild type and makes it possible to explore if a certain mutation increases or decreases affinity to the target. This library of phage can bind to molecules of the immobilized target and then the ones that didn't succeeded in binding get washed away. The remaining bound phages are used to infect bacteria that produce other phages carrying the same genetic information (amplification phase). And the experiment can be repeated from the beginning.

Just after the amplification phase the bound phages can be sequenced obtaining the number of bound phages displaying a particular sequence. A schematic representation reprinted from [6] can be found in fig. 3.1.

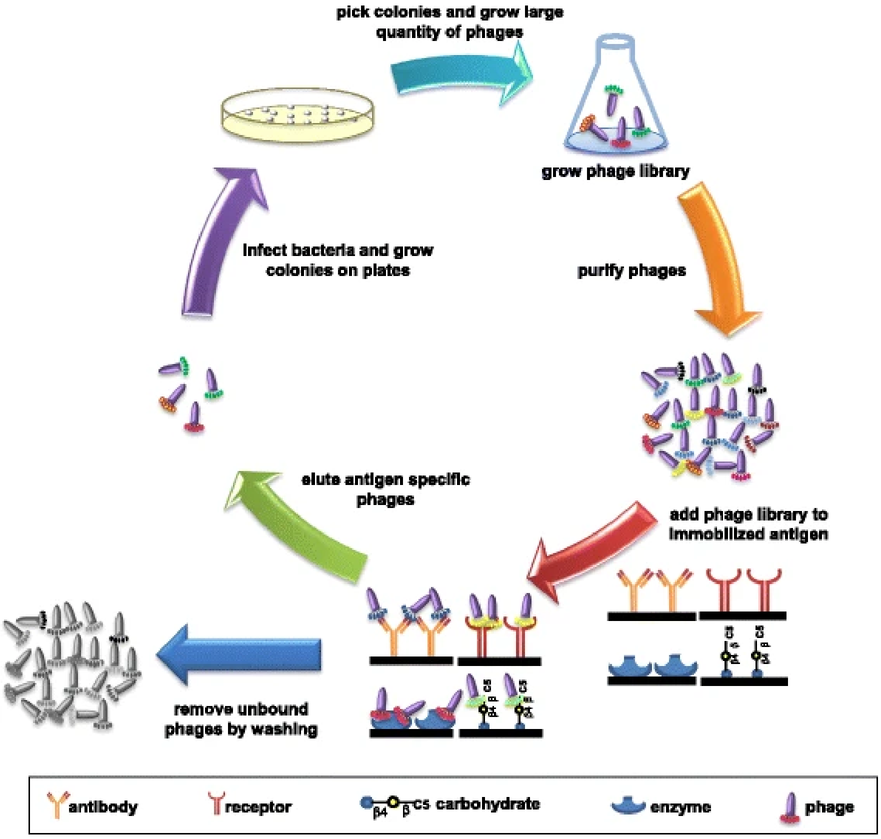

pick colonies and grow large quantity of phages
grow phage library
purify phages
add phage library to immobilized antigen
remove unbound phages by washing
elute antigen specific phages
infect bacteria and grow colonies on plates
antibody
receptor
β4 β CS carbohydrate
enzyme
phage

Figure 3.1: Schematic representation of the various steps of a phage display experiment. Reprinted from [6]

.

3.2 Machine Learning technique

Before the work exposed in [7] this problem was usually tackled with standard supervised learning techniques: a proxy was computed for the affinity of a sequence from the reads of the various rounds, and then a standard machine learning methods solved the regression.
The unsupervised method exposed in this work models the binding process as a probabilistic one writing down the binding probability of each sequence

depending on its specific composition through some parameters. Using the sequences reads at various rounds, a Likelihood function can be written in terms of these parameters and it can be maximized to find the best set of parameters that describes the experiment. And because of the specificity of these parameters they can be used to infer the affinity of sequences that were not observe in the experiment.

3.3 Statistical Model

Consider a library of sequences composed by S different variants of a wild type. What is observed from phage display experiments is $N_s(t)$, the number of phages displaying sequence s at the beginning of round t ($t = 1, ..., T$) where $N_s(1)$ represents the initial abundance of variant s in the library used for the experiment.

A single round con be modelled as a combination of two stochastic processes: a first phase of selection during which the phages have the possibility to bind to the target molecules and then the unbound phages are washed away. After this, there is a second phase during which the phages that survive the wash are used to infect bacteria that are used to regrow the library until it reaches the starting size.

The selection phase is modelled using a binomial distribution: every variant can bind to the target with probability p_s so the probability to find $n_s(t)$ phages carrying sequence s bound to target during round t is given by

$$P(n_s(t)|N_s(t), p_s) = \binom{N_s(t)}{n_s(t)} p_s^{n_s(t)} (1 - p_s)^{N_s(t) - n_s(t)} \tag{3.1}$$

Since it is reasonable to consider different rounds as statistically independent, the probability of the selected variants through the entire experiment can be written as

$$P(\{n_s(t)\}_{t,s}|\{N_s(t)\}_{t,s}, \{p_s\}_s) = \prod_{t=1}^{T-1} \prod_s \binom{N_s(t)}{n_s(t)} p_s^{n_s(t)} (1 - p_s)^{N_s(t) - n_s(t)} \tag{3.2}$$

where the notation $n_s(t)_{st}$ stands the set of all $n_s(t)$ for $s = 1, ..., S$ and $t = 1, ...T - 1$ (the values of t included in this set will be obvious from the context, for instance in this case $t = T$ is not included because $N_s(T)$ is determined by the amplification phase of round $T - 1$ and round T has no

selection phase because it represents the end of the experiment.
Instead, the amplification phase can be modelled with a multi-nomial distribution with a probability of producing a new phage with variant s given by the fraction of phages carrying sequence s that are used to infect bacteria, so

$$P(\{N_s(t+1)\}_s|\{n_s(t)\}_s) = \frac{N_{tot}!}{\prod_s N_s(t+1)!} \prod_s \left(\frac{n_s(t)}{n_{tot}(t)}\right)^{N_s(t+1)} \tag{3.3}$$

where $N_{tot} = \sum_s N_s(t)$ and it is independent of t because it has been assumed that the amplification phase restore the initial abundance of each variants, and $n_{tot}(t) = \sum_s n_s(t)$ is the total number of bound phages at round t.
Since it can be assumed that the rounds of the experiment are independent one from another it is possible to write the Likelihood for the whole experiment as

$$\begin{aligned} &P(\{N_s(t)\}_{s,t}, \{n_s(t)\}_{s,t}|\{p_s\}_s, N_{tot}) = \\ &\qquad \underbrace{P(\{n_s(t)\}_{t,s}|\{N_s(t)\}_{t,s}, \{p_s\}_s)}_{\text{selection phases}} \underbrace{\prod_{t=1}^{T} P(\{N_s(t+1)\}_s|\{n_s(t)\}_s)}_{\text{amplification phases}} \end{aligned} \tag{3.4}$$

where N_{tot} must be considered given because it can be computed from the starting library. Since during the experiment it is not possible to observe $n_s(t)$ they can be traced out by marginalizing the and write the Likelihood as function of only the variables $N_s(t)$ as

$$\begin{aligned} &P(\{N_s(t)\}_{s,t}|\{p_s\}_s, N_{tot}) = \\ &\qquad \sum_{\{n_s(t)\}_{s,t}} \underbrace{P(\{n_s(t)\}_{t,s}|\{N_s(t)\}_{t,s}, \{p_s\}_s)}_{\text{selection phases}} \underbrace{\prod_{t=1}^{T} P(\{N_s(t+1)\}_s|\{n_s(t)\}_s)}_{\text{amplification phases}} \end{aligned} \tag{3.5}$$

At this point the crucial step is how to compute 3.5, and it is clear that it is not possible to perform it analytically: one must resort to some kind of approximation.
The first approximation that one can do is to assume that $n_s(t)$ assumes the value given by the average of the binomial distribution that determines it, i.e. $n_s(t) \approx p_s N_s(t)$. This approximation is called *deterministic binding approximation* and by using it and taking the logarithm, the log-likelihood

to be maximized becomes $\mathcal{L} = \sum_{s,t} \mathcal{L}_{s,t}$, with $\mathcal{L}_{s,t}$ given by

$$\mathcal{L}_{s,t} = N_s(t+1) \log \frac{p_s N_s(t)}{\sum_\sigma p_\sigma N_\sigma(t)} \tag{3.6}$$

This approximation has been used to obtain the results exposed in [7]. The other possible approximation can be introduced in the amplification phase: it will be no longer a stochastic step but a deterministic one and the bound phages get amplified by a factor $\lambda(t)$ independent of the variants, such that $N_s(t+1) = \lambda(t) n_s(t)$. So, at the cost of introducing $T - 1$ new parameters it is possible to eliminate the variables $n_s(t)$ by using the change of variables $n_s(t) = N_s(t+1)/\lambda(t)$. Therefore 3.1 becomes

$$\begin{aligned} P(N_s(t+1)|N_s(t), p_s, \lambda(t)) = \\ \frac{1}{\lambda(t)} \binom{N_s(t)}{N_s(t+1)/\lambda(t)} p_s^{N_s(t+1)/\lambda(t)} (1-p_s)^{N_s(t) - N_s(t+1)/\lambda(t)} \end{aligned} \tag{3.7}$$

Multiplying it for all variants and all rounds to get the likelihood of the data obtained in the experiment and taking the logarithm one gets

$$\mathcal{L}_{s,t} = -\log \lambda(t) + \log \binom{N_s(t)}{\frac{N_s(t+1)}{\lambda(t)}} + \frac{N_s(t+1)}{\lambda(t)} \log \frac{p_s}{1-p_s} + N_s(t) \log(1-p_s) \tag{3.8}$$

This approximation (it will be referred to as *uniform amplification factor approximation*) will be used in the present work and the results will be compared with the ones obtained employing deterministic binding approximation. Note that in this case the condition that amplification restores the initial size of the library is not ensured but it is not a problem because if the variables $N_s(t+1)$ are rescaled, such rescaling produces only a constant term that can be neglected in the maximization operation, as long as one is interested only in the model parameters concerning the binding probabilities, because it doesn't depend on any of the parameters of the model. And because of this "rescaling invariance" the variables $N_s(t+1)$ can be considered as continuous ones and the logarithm of the binomial coefficient appearing in 3.8 can be expressed using Stirling approximation

$$\log \binom{n}{k} \approx n \log n - k \log k - (n-k) \log(n-k) \tag{3.9}$$

3.3.1 Specificity

What is left to model is the specificity of the binding probabilities p_s in terms of the primary structure of the sequences. This can be done assuming that we can model each sequence as a two-states (*bound/unbound*) system at equilibrium, where the unbound state corresponds to the reference level for energy ($H_s = 0$) and the energy of the bound state is given by the Hamiltonian of a Potts Model which takes into account single and double occurrences of residues in the sequence chain

$$H_s = -\sum_{i<j} J_{ij}(s_i, s_j) - \sum_i h_i(s_i) \tag{3.10}$$

It is well known that such a system follows the Fermi-Dirac statistics and p_s is given by

$$p_s = \frac{1}{1 + \exp(H_s - \mu)} \tag{3.11}$$

where μ is the chemical potential and it depends on the target concentration. Inserting 3.11 in 3.8 one gets the complete expression of the likelihood to maximize as function of the model parameters $J_{ij}(s_i, s_j), h_i(s_i), \mu, \lambda(t)$ for every i, j, s_i, s_j, t.

The expression of p_s in eq. 3.11, as usually done [8], easily follows from a grandcanonical ensamble description inserting a binary variable $\sigma \in \{0,1\}$ that describes these two thermodynamical states ($\sigma = 0$ corresponds to the unbound state and $\sigma = 1$ corresponds to the bounded state). Then the grandcanonical partition function $\mathcal{Z}$ reads, setting the inverse temperature $\beta = 1$ since it plays no role in this description and can be considered as being incorporated to the parameters of the Hamiltonian:

$$\mathcal{Z} = \sum_{\sigma\in\{0,1\}} e^{-(H_s-\mu)\sigma} = 1 + e^{-(H_s-\mu)} \tag{3.12}$$

and the expression in eq. 3.11 is the normalized Boltzmann weigth

$$p_s = \frac{e^{-(H_s-\mu)}}{1 + e^{-(H_s-\mu)}} = \frac{1}{1 + \exp(H_s - \mu)} \tag{3.13}$$

3.3.2 Likelihood maximization

When performing the maximization process some conditions must be taken into account.

First of all the condition $n_s(t) \leq N_s(t)$ it must be always satisfied for every variant at each time, and in the uniform amplification factor approximation it means that $\lambda(t) \geq \frac{N_s(t+1)}{N_s(t)} \quad \forall s$, therefore

$$\lambda(t) \geq \max_s \left\{ \frac{N_s(t+1)}{N_s(t)} \right\} \tag{3.14}$$

In principle equality in formula 3.14 would be admissible for the model because the term $(N_s(t) - N_s(t+1)/\lambda(t)) \log(N_s(t) - N_s(t+1)/\lambda(t))$ would vanish, so it remains finite, if thought as a limit of the type $\lim_{x \to 0} x \log x$; But when performing maximization the algorithm employed to search for the maximum uses the gradient of the likelihood to move from point to point, and computing the gradient of the likelihood with respect to $\lambda(t)$ one gets

$$\frac{\partial \mathcal{L}}{\partial \lambda(t)} \approx \\ \sum_s \frac{N_s(t+1)}{\lambda(t)^2} \left\{ \log \frac{N_s(t+1)}{\lambda(t)} - \log \left(N_s(t) - \frac{N_s(t+1}{\lambda(t)} \right) + H_s - \mu \right\} - \frac{S}{\lambda(t)} \tag{3.15}$$

where the symbol "≈" has been used because of the Stirling approximation. From this expression one can easily see that the second term of this expression becomes singular when in the condition 3.14 the equality holds. For this reason when performing numerical maximization this case must be excluded and 3.14 becomes

$$\lambda(t) > \max_s \left\{ \frac{N_s(t+1)}{N_s(t)} \right\} \tag{3.16}$$

and this condition has been provided by adding a small value (δ) at the border of the constraint

$$\lambda(t) \geq \max_s \left\{ \frac{N_s(t+1)}{N_s(t)} \right\} + \delta \tag{3.17}$$

Besides, something more can be said about the location of the maximum of the log-likelihood as function of $\lambda(t)$. Since the optimization over $\lambda(t)$ is constrained, the maximum could be attained at the border without the gradient to be zero. By rewriting equation 3.8 in a clearer way by inserting both the functional form of H_s and the Stirling approximation for the binomial coefficient one gets

$$\mathcal{L} = \sum_t \sum_s \Big[\frac{N_s(t+1)}{\lambda(t)}(\mu - H_s) - \frac{N_s(t+1)}{\lambda(t)} \log \frac{N_s(t+1)}{\lambda(t)} \\ - \Big(N_s(t) - \frac{N_s(t+1)}{\lambda(t)}\Big) \log \Big(N_s(t) - \frac{N_s(t+1)}{\lambda(t)}\Big) - \log \lambda(t) \Big] \quad (3.18)$$

When $\lambda(t)$ reaches the lower border it means that there is (at least) a sequence such that $N_s(t) - \frac{N_s(t+1)}{\lambda(t)} = 0$. In this case the log-likelihood expression in equation 3.18 remains finite. The gradient expression in equation 3.15 instead, diverges to $+\infty$ and this means that the log-likelihood as function of $\lambda(t)$ at the border, incrases with a vertical asymptote.

Studying the limit of 3.18 for $\lambda(t) \to +\infty$ one can see that the term

$$\frac{N_s(t+1)}{\lambda(t)} \log \frac{N_s(t+1)}{\lambda(t)}$$

approaches zero and the whole expression tends to $-\infty$. If the maximum is not attained at the border it must be a stationary point, implying that $\frac{\partial \mathcal{L}}{\partial \lambda(t)} = 0$, which corresponds to

$$\frac{\partial \mathcal{L}}{\partial \lambda(t)} \approx \\ \sum_s \frac{N_s(t+1)}{\lambda(t)^2} \Big\{ \log \frac{N_s(t+1)}{\lambda(t)} - \log \Big(N_s(t) - \frac{N_s(t+1}{\lambda(t)}\Big) + H_s - \mu \Big\} \\ - \frac{S}{\lambda(t)} = 0$$

$$\sum_s N_s(t+1) \log \Big(\frac{N_s(t)\lambda(t)}{N_s(t+1)} - 1\Big) + S\lambda(t) = \sum_s N_s(t+1)(H_s - \mu) = 0 \quad (3.19)$$

Since the l.h.s. of 3.19 is strictly increasing with $\lambda(t)$ the solution of equation 3.19 is unique. Considering what has just been stated before and together with the fact that the solution of $\frac{\partial \mathcal{L}}{\partial \lambda(t)} = 0$ is unique, one can conclude that the log-likelihood as function of $\lambda(t)$ is concave and the maximum can't be attained at the border.

In summary the likelihood maximization consists in a constrained optimization of the sum of all contributions given by eq. 3.8 with respect to the model parameters, namely $J_{ij}(a,b), h_i(a), \mu, \lambda(t)$ for all a, b in the vocabulary, i, j from 1 to L and t from 0 to $T-1$.

3.4 Validation of the inference

A major issue in this approach is that there is no quantity that can be determined experimentally and can be used as proxy to asses the validity of the learnt model. Nevertheless such a quantity can be computed from the sequences reads [9]: the only information provided by experiments.
The selectivity of a sequence, which is defined as the ratio of reads at two consecutive rounds, can be interpreted as a measure of the affinity of the sequence to the target. Considering the ideal case in which both the approximations hold

$$N_s(t+1) = \lambda(t) n_s(t) \quad \text{uniform amplification} \tag{3.20}$$

$$n_s(t) = p_s N_s(t) \quad \text{deterministic binding} \tag{3.21}$$

one gets that the following condition should hold

$$\frac{N_s(t+1)}{N_s(t)} = \lambda(t) p_s \tag{3.22}$$

Taking logarithms and inserting a term $\epsilon_s(t)$ that represents the measurement error one gets

$$\log N_s(t+1) - \log N_s(t) = \theta_s + \alpha(t) + \epsilon_s(t) \tag{3.23}$$

where θ_s is the (empirical) log-selectivity, $\alpha(t)$ is $\log \lambda(t)$ and they can be determined with a least squares fit where the data are the sequencing reads $N_s(t)$ assuming $1/N_s(t)$ as variance for them, as if they were sampled from a Poisson's distribution as described in [9] and used in [7].
The results from this fit provides also an estimate for the error that must be associated to the empirical selectivities θ_s.
Once these quantities θ_s hve been determined, the accuracy of the learning is assessed by computing the correlation between the energy E_s and the log-selectivity θ_s of the sequences of the dataset. Two measures of correlation will be considered in this work: Pearson coefficient and Spearman correlation.

3.5 Datasets

This short section contains a description of the datasets analyzed in this work [10] [11] [12]: relevant properties are reported in table 3.1. The second column reports the length of the mutated sequence, the third one reports the number of rounds that have been sequenced during the experiment, the third column reports the number of sequences whose counts are are consistent with the statistical model provided in section 3.3, that is to say that all sequences that at a certain round have zero reads and then reappear in a following, due to experimental error in the sequencing step, round have been discarded from all datasets. The reason is that if $N_s(t) = 0$ one must have $N_s(t + 1) = 0$, because $N_s(t) = 0$ forces $n_s(t)$ to be zero. Finally, the last column reports the ratio between the number of training sequences (training-test split of 80%-20% is always considered here) and the number of parameters to learn. The first dataset is from a 2016 study by Boyer et al. that involves antibodies.

Reference	Variant length	Rounds	Valid sequences	Sequences per parameter
Boyer et al.	4	2	23572	≈7.6
Fowler et al.	25	2	478733	≈3.2
Wu et al.	4	1	158447	≈51.1

Table 3.1: Datasets description

Antibodies have a very complex protein structure but all of them share some common features (see fig. 3.2. They have a particular particular "Y" shape where the two upper chains are highly variable from one to another and they are the part of the molecule which can bind to the antigen. These chains, in turn, are divided in an *heavy chain* ($\mathrm{V_H}$) and a *light chain* ($\mathrm{V_L}$). To construct the dataset they have focused in inducing mutation in 4 particular sites in the $\mathrm{V_H}$ chain, reported as CDR3 in fig. 3.3, producing all possible combinations for these sites. As target molecule they have used a fragment of hairpin DNA.

The second dataset is from a 2010 study by Fowler et al.. It has been performed using, as variants, mutated WW domain of human YAP65, a protein involved in transcription regulation (see fig. 3.4), and its peptide ligand as target.

The third dataset is from a 2016 study by Wu et al.. They have induced mutations in four particular sites of GB1 protein to construct the library of variants (see fig. 3.5). And they have used the *fragment crystallizable region* (Fc region) of the IgG antibody as target that can interact with protein G.

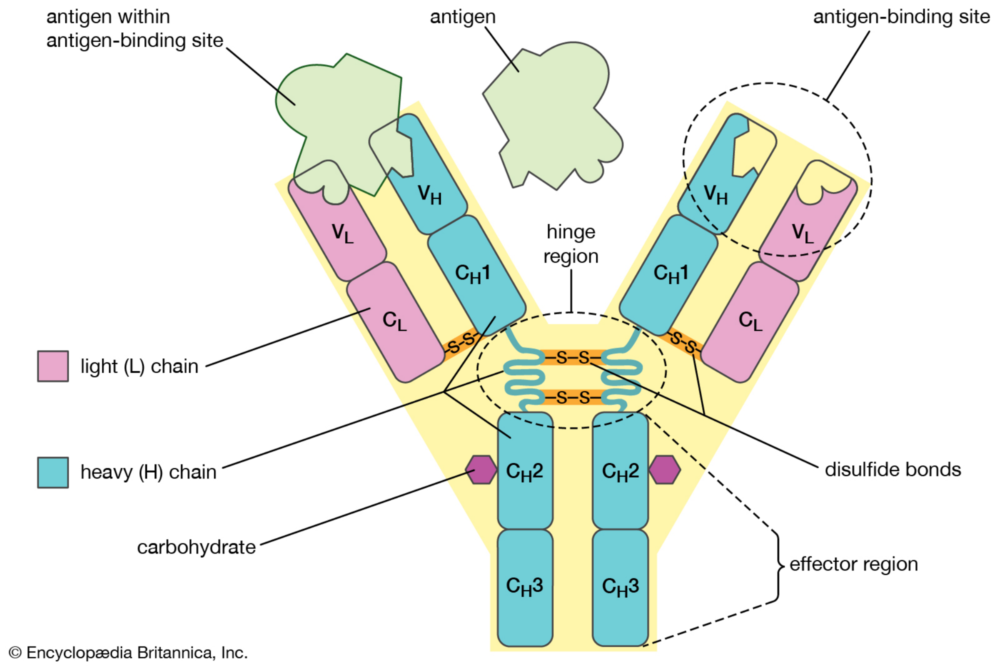


Figure 3.2: Antibody's structure. Reprinted from "Britannica, The Editors of Encyclopaedia. "Antibody". Encyclopedia Britannica, 27 May. 2020, https://www.britannica.com/science/antibody."

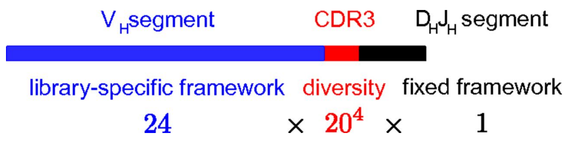


Figure 3.3: Library design for the experiment by Boyer et al.. In red is reported the variant region (CDR3) and all possible combinations of residues for these sites have been included in the library. Reprinted from [10]

3.6 Analysis and Results

3.6.1 Preprocessing

A usual preprocessing step that is taken in these cases is that of filtering out sequences with low counts during rounds (all but the last one). The reason for this is that these sequences introduce overfitting in the learned model because of the high fluctuations in the fraction of bound phages carrying these sequences. This procedure usually results in filtering out sequences that

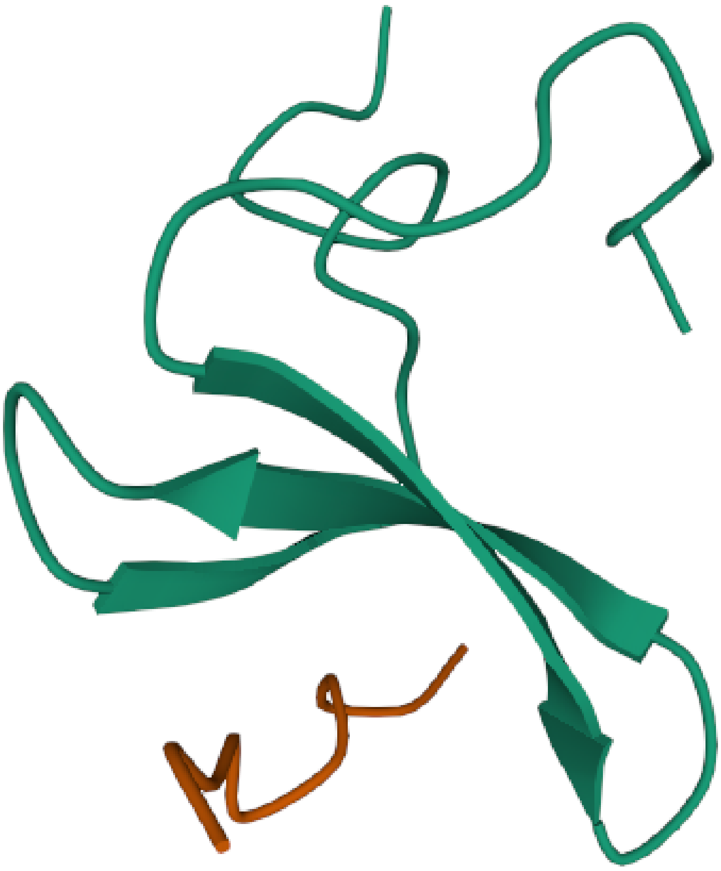


Figure 3.4: Protein structure of human YAP65, WW domain is coloured in red. Reprinted from Protein Data Bank, identifier 1JMQ.

have an high error on the log-selectivity θ_s from the least square fit (see figure 3.6). Particular attention must be paid when performing this step. When the length of the mutated sequence grows the number of parameters involved in the statistical model increase dramatically. When filtering sequences one should always obtain is a situation in which the number of sequences in the training set is much grater than the number of parameters. When this condition is not met, an accurate and robust learning is impossible. In fact this preprocessing step has not been performed because the only dataset that would have allowed this step is the one from Wu et al. but as will be shown later the performance is so good that this step would have introduced minimal changes in the performance.

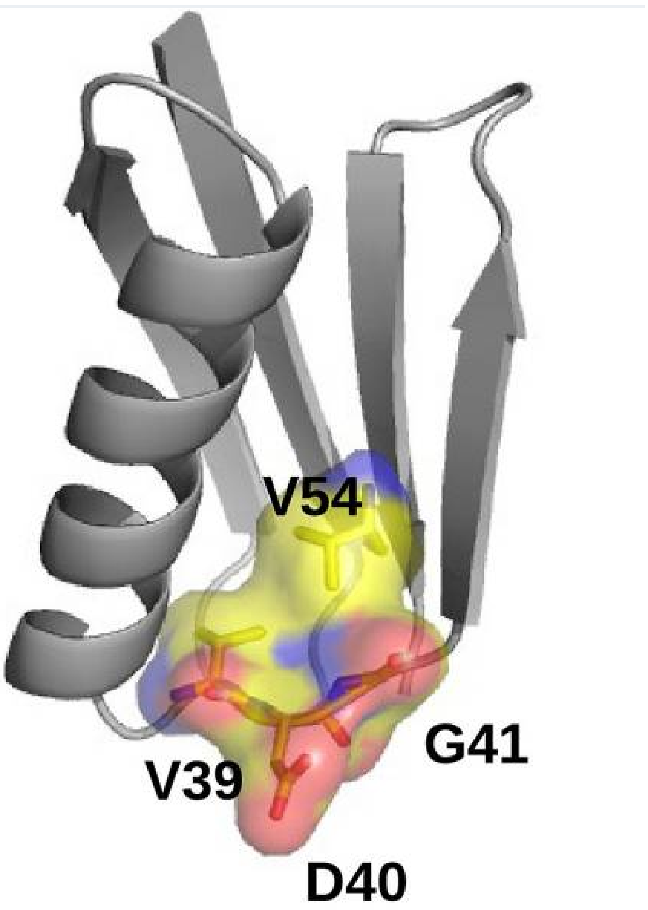
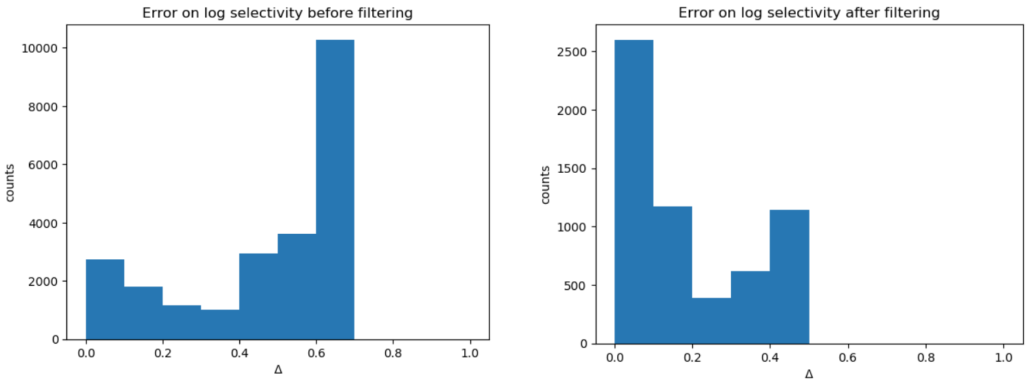


Figure 3.5: GB1 protein's structure. The four mutated sites are labelled and highlighted in the picture. Reprinted from [12]

Because these low-counts sequences has been kept for the learning a regularization has been introduced in the log likelihood. In particular an L2-norm regularization has been used, resulting in a total likelihood

$$\mathcal{L}_{total} = \mathcal{L} - \eta_J \sum_{a,b,i<j} J_{ij}(a,b) - \eta_h \sum_{i,a} h_i(a) - \eta_\mu \mu^2 \tag{3.24}$$

In equation 3.24 the various couplings, fields and chemical potential have different number of terms. So the parameters η_J and η_h have been divided by the number of parameters of these two types, in order to give the same contribution to the total log likelihood. When a large number of low-counts sequence is used in the learning, the values of these parameters must be increased accordingly. The following and last preprocessing step is the



(a) Reads before filtering out low counts sequences

(b) Reads after filtering out sequences with less than 3 counts at the beginning of the experiment

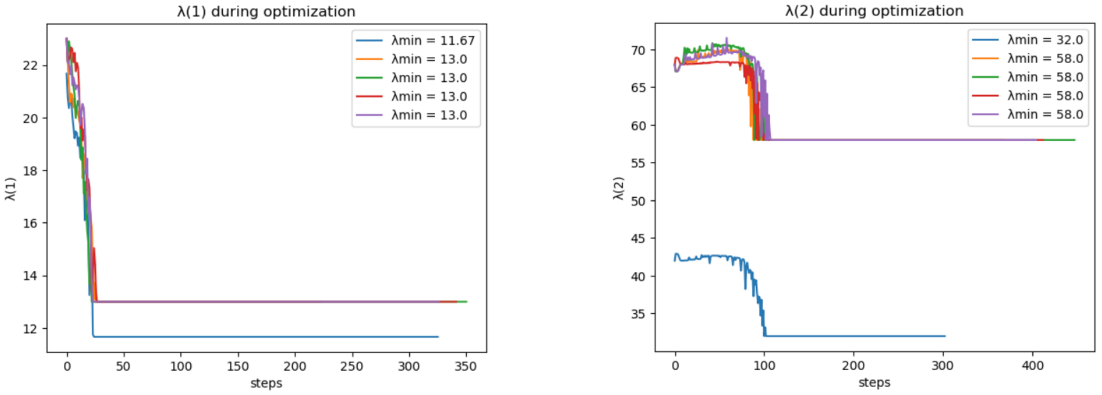
Figure 3.6: Reads before and after filtering out sequences with low counts for the Boyer et al. dataset

addition of pseudocounts: 0.5 for all sequences. In particular the same has been done before computing the empirical log-selectivities as already discussed in section 3.4, as reported in [7].

3.6.2 Learning

The learning has been performed through a 5-fold crossvalidation, as done in [7], for multiple reasons: to compare results, to make up for the small number of training sequences and, finally, to take into account the fact that train-test split is carried out randomly.
The non-linear optimization has been performed by a numerical routine using the Method of Moving Asymptotes[1] because this algorithm has been observed to be not very sensitive to the tolerance of the objective function. Some snapshots of the learning process of the amplification factor are present in figure 3.7. It can be seen that at the end of the learning, convergence is reached and this means that $\lambda(t)$ sets correctly on the unique maximum and does not fluctuate. This is expected as said before, because the solution of

[1] The specific package used for the numerical optimization is NLopt, it implements non-linear optimization algorithms. Reference is available at `https://nlopt.readthedocs.io/en/latest/`

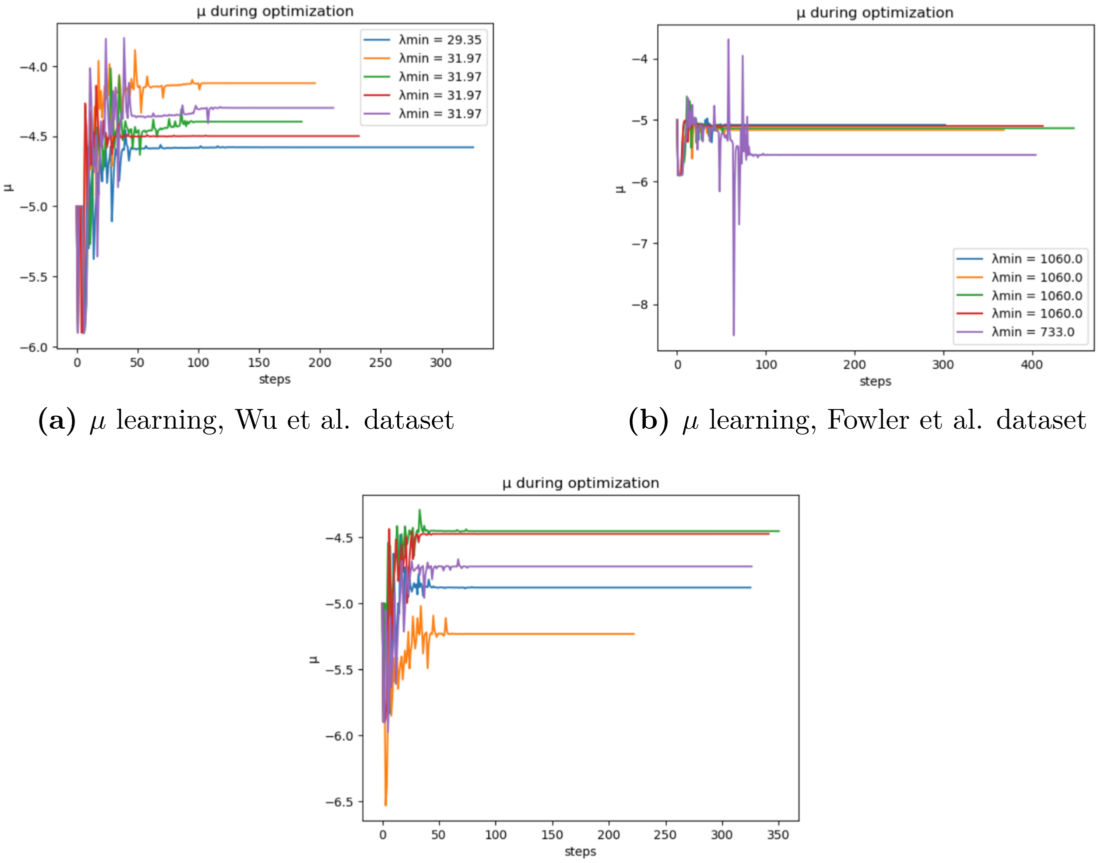


(a) $\lambda(1)$ learning, Boyer et al. dataset

(b) $\lambda(2)$ learning, Fowler et al. dataset

Figure 3.7: Learning of the parameter $\lambda(t)$ in some cases.

equation 3.19 is unique.
One can notice that, in some cases, like the one in panel 3.7b the lower bound for $\lambda(t)$'s at different rounds can differ considerably (about 50% in this case). For this reason is crucial to repeat the learning as many times as possible to mitigate this effect produced by a random instance of the training set.
What is reported in figure 3.7 can seem to be in contrast with the fact that the maximum can't be attained at the border as already stated in 3.3.2, but the explanation resides in the fact that to exclude the value at the border a small quantity δ has been introduced in 3.17, and since at the border the function has a vertical asymptote it can increase a lot within a very small displacement from the border.
Other relevant quantities can be observed during the learning process, for instance the chemical potential reported in figure 3.8. From it, it can be seen that in all cases μ reaches a pretty highly negative value. Nevertheless this value should be compared with the energies of the sequences, and the rare-binding regime holds only if the energy of the sequences are much larger than the chemical potential. The graph in figure 3.9 has been obtained by learning the model parameters from an instance of train set and then using these parameter to estimate the energies on the train set. From this figure one can see that the inferred probabilities start to differ from the Maxwell-Boltzmann when energy (μ included) is below 2.0. One can count the fraction of sequences for which energy doesn't exceed this value among all the sequences, this fraction is $\approx 98\%$ for both training and test set. One can



(a) μ learning, Wu et al. dataset

(b) μ learning, Fowler et al. dataset

(c) μ learning, Boyer et al. dataset

Figure 3.8: Snapshots of the learning for μ parameter in some cases.

conclude that the fraction of sequences outside of the Maxwell-Boltzmann regime can be negligible respect to the ones in this regime. In this cases the binding probability can be approximated as

$$p_s \approx \exp(-(H_s - \mu)))$$

Obviously this condition should be checked from time to time.

3.6.3 Validation

After all these considerations one can assess the validity of the learning as exposed in 3.4. Figure 3.10 shows the correlation between energies and log-selectivities of the sequences as function of fraction of data retained:

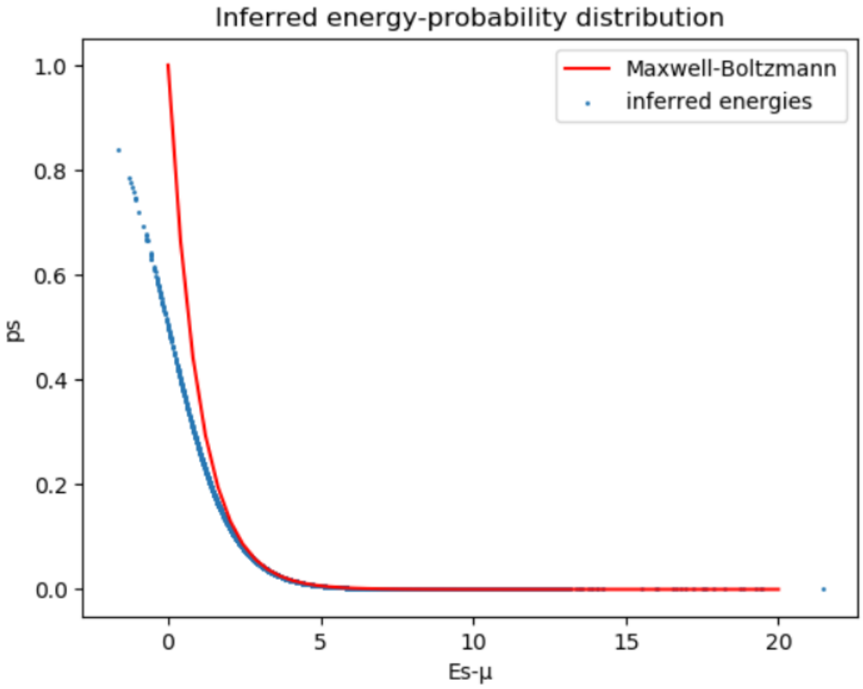
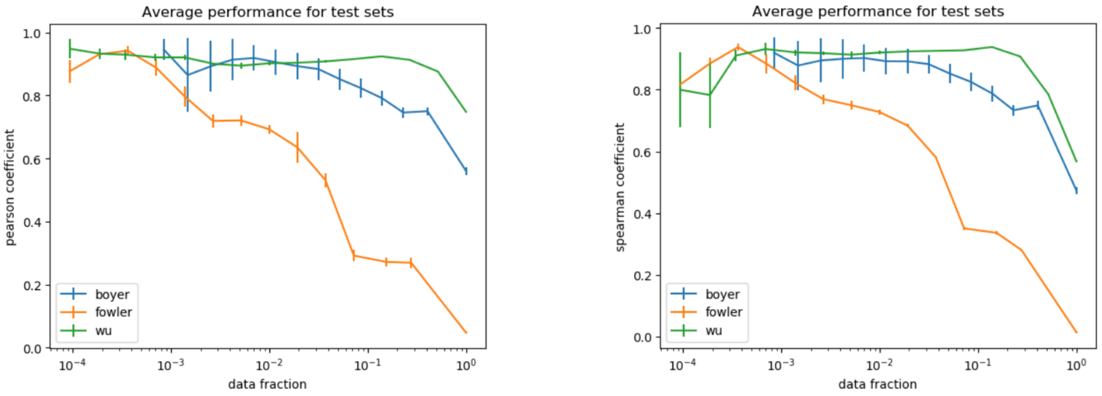


Figure 3.9: Comparison of the inferred probabilities on a test set with the Maxwell-Boltzmann distribution. Form Wu et al. dataset.

correlation has been computed multiple times, each time discarding sequences with highest error on the log-selectivity. It can be seen that by keeping sequences with smaller values of errors on selectivities the correlation increases and it reaches pretty high values of correlation. The absence of overfitting can be assessed by inspecting the graphs in figure 3.11. Figure 3.11b shows that the curves relative to training set and test set have a very small discrepancy indicating that the model can generalize well in predicting the selectivity of new samples.

By looking at table 3.1 one can see that the dataset from Fowler et al. has the smallest ratio between training samples and model parameters, and this is the case which is most sensible to overfitting. But by looking at 3.11a it can be seen that this is prevented by the increasing of the regularization as described before. The ultimate validation that one can think about to assess the validity of the model in extreme learning situation: to learn the model parameters on sequences with the lowest selectivity and testing the results on sequences with high selectivity. One can do this to asses if the model is capable to catch relevant single and double occurrences that are really linked



(a) Pearson correlation

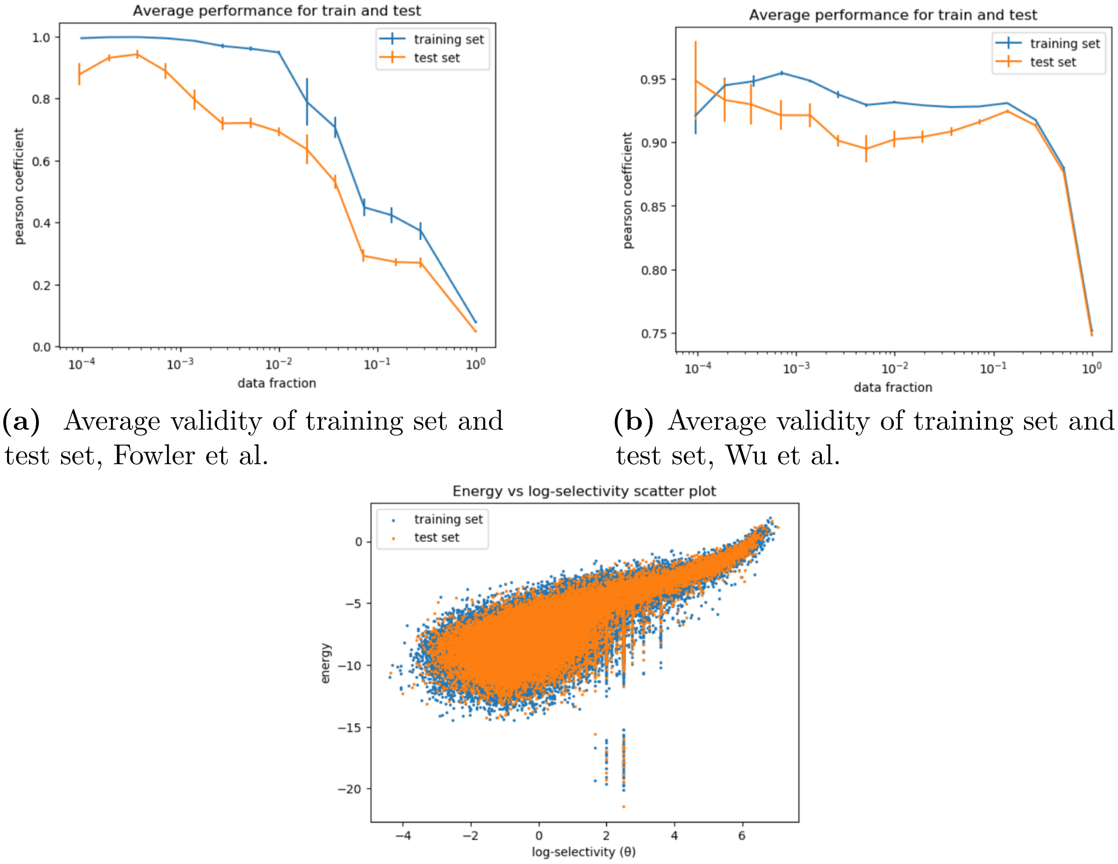
(b) Spearman correlation

Figure 3.10: Correlation between log-selectivity and energies as function of the fraction of retained data for all datasets. Sequences with high errors on empirical-log-selectivity are discarded

to the affinity with the target.
This test has been conducted on Wu et al. dataset because it showed the most robust and valid learning due to the high number of samples. Figure 3.12 shows an histogram with the values of log-selectivity for the Wu et al. dataset and shows how they have been divided into training set and test set. After the learning has been carried out the validation of the model shows some interesting results, reported in figure 3.13. In the same way it happens when performing standard cross-validation (3.10) the correlation succeeds in remaining high even for large fraction on test set (where sequences with an high error on log-selectivity are kept) but there is a sudden drop when the test set gets decimated. Even though a some overfitting seems to be present, the validity of the model on the test set remains constant for a very wide range of different data fraction. This can mean that if one looks only at sequences with high selectivity (usually they are the most relevant one during experiments) the learning shows to be robust and stable.

3.7 Drawbacks of the model and future work

Although, assessing the validity of this model, quite remarkable results have been obtained there are still some drawbacks of the model which can possibly open the way for future work.



(a) Average validity of training set and test set, Fowler et al.

(b) Average validity of training set and test set, Wu et al.

(c) Scatter plot energy vs empirical log selectivity

Figure 3.11: Correlation between log-selectivity and energies as function of the fraction of retained data for all datasets, computed on training set and test set of the same dataset. Dataset from Wu et al.

The main issue is that of separating the effect of binding probabilities from the effect of the amplification factor. In going from a certain number of reads $N_s(t)$ to a number of reads $N_s(t+1)$ at the following rounds many combinations of binding and amplification processes can describe the data. For instance, just as a clarifying example: if $N_s(t) = 6$ and $N_s(t+1) = 10$ both $(p_s = 1/3, \lambda(t) = 5)$ and $(p_s = 1/6, \lambda(t) = 10)$ can fit these data, as well as many other combinations of p_s and $\lambda(t)$. Nevertheless, up to a certain point, the learning tends to prefer those combination for wich p_s is not too close to 0 or 1 due to the effect of the regularization on the model parameters

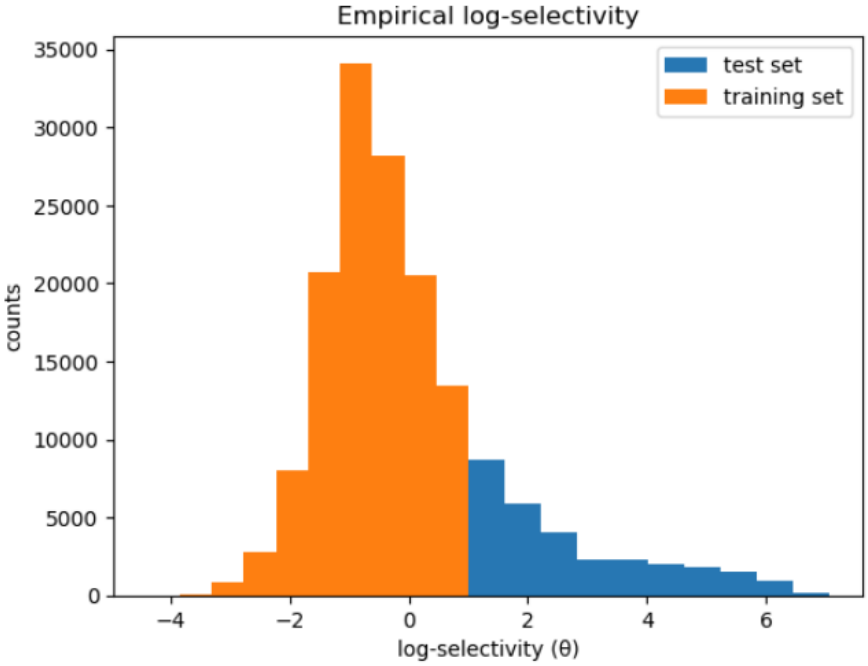
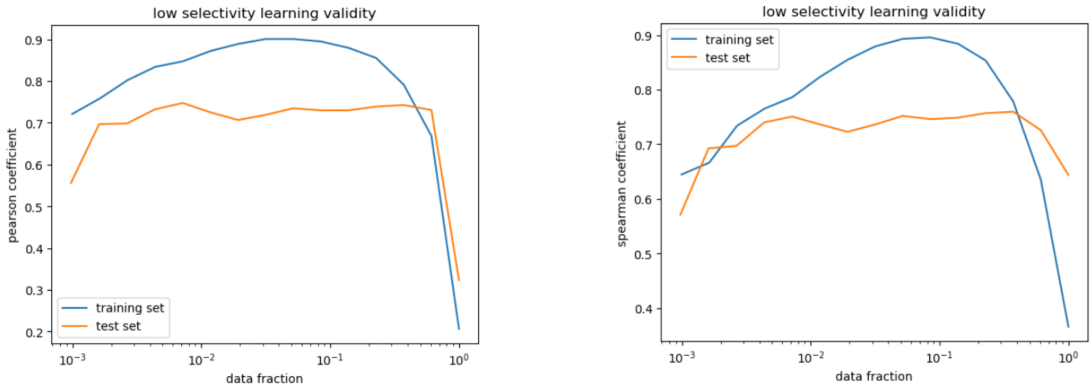


Figure 3.12: Histogram of empirical log-selectivity for Wu et al. dataset. The sequences which have been put in the training set and in the test set has been coloured differently. Train/test proportion is still around 4/5.



(a) Pearson coefficient

(b) Spearman coefficient

Figure 3.13: Validation of the model learned on low-selectivity sequences and tested on the high-selectivity ones.

$J_{ij}(a,b), h_i(a)$ and μ but still no regularization has been introduced on $\lambda(t)$ because it is not clear the way it should be done.
Another drawback concerning the parameter $\lambda(t)$ is the strong dependence on the training set instance. The values of $\lambda(t)$ can be interpreted as the average infected bacteria during the amplification phase at round t, it is expected to be independent of the particular sequence and to be a property of the phages. The fact that a particular training set sets a lower bound for $\lambda(t)$ can make the inferred $\lambda(t)$ differ considerably from a learning to another (this situation is depicted in figure 3.7b). This in turn can imply an high error to be associated with this parameter.
Future work can certainly aim to expanding and enriching the model. This particular model considers only two states of the sequence, it can be found either *bound* or *not bound* to the target. But in practice one could distinguish this last case in two possible sub-cases: one in which the sequence is folded and active to perform its task and bind to the target, and a second one in which the sequence is in an unfolded state and in any case it would be impossible for it to bind to the target. In the first case one would associate the event of *non binding* to a low affinity of that particular sequence with the target, in the other case one associate the *non binding* the the fact that binding has been prevented by mechanical reasons; this can avoid to introduce a certain amount of bias in the results. This situation can be modelled with a three-states system instead of two; one associated to the state *unfolded*, another one to the state *folded but not bound* and a third one to the state *folded and bound*. This model can be explored but the main issue is that, in contrary to this two-states model, the likelihood maximization is not so trivial because one would obtain a non concave function and a way to skip local maxima must be employed.

3.8 Conclusions

The positive validation showed in fig. 3.10 suggests that when the empirical estimate of the selectivity is reliable we can observe a large correlation with the one inferred on the test set marking a good agreement between experimental results and this machine learning approach (the choice of the metric is a minor issue in this case since they show the same behaviour). And when the experimental dataset is consistent and it allows for a robust and reliable learning these results don't depend crucially on the fraction of data retained, this is a clear sign of the fact that the statistical model is capable to capture the essential features that makes the sequence bind to the target.

This particular approximation is also capable to give insights into the value of the amplification factor $\lambda(t)$. Because of the good performance of the learning one could conclude that considering this quantity as independent of the particular sequence bound to the target, and this is precisely what is experimentally expected. A result, based a probabilistic analysis, that can be useful also on the experimental side.

These result are relevant as long as the system is modelled as a two-state one, but more work in due when one wants to improve the accuracy of this model and its capability to capture capture the features of the experimental result by introducing additional states: in this case some implementation issues arises as described in sec. 3.7. But the good results obtained since now outlines the validity of these approaches based on statics as valid contributors to practical issues such as protein design, antibodies engineering and further developments in the phage display experimental technique and its several applications.

Bibliography

[1] 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 pp. 2, 3).

[2] 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 p. 4).

[3] David J. C. MacKay. *Information Theory, Inference Learning Algorithms*. USA: Cambridge University Press, 2002. ISBN: 0521642981 (cit. on p. 7).

[4] 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://doi.org/10.1088/1361-6633/aa9965 (cit. on pp. 8, 15).

[5] 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». In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72. ISSN: 0027-8424. DOI: 10.1073/pnas.0805923106. eprint: https://www.pnas.org/content/106/1/67.full.pdf. URL: https://www.pnas.org/content/106/1/67 (cit. on p. 8).

[6] Wu, Chien-Hsun, Liu, I-Ju, Lu, Ruei-Min, Wu, and Han-Chung. «Advancement and applications of peptide phage display technology in biomedical science». In: *Journal of Biomedical Science* 23.8 (2016). DOI: https://doi.org/10.1186/s12929-016-0223-x (cit. on p. 19).

[7] Jorge Fernandez-de-Cossio-Diaz, Guido Uguzzoni, and Andrea Pagnani. «Unsupervised Inference of Protein Fitness Landscape from Deep Mutational Scan». In: *Molecular Biology and Evolution* 38.1 (Aug. 2020), pp. 318–328. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa204. eprint: https://academic.oup.com/mbe/article-pdf/38/1/318/35389203/msaa204.pdf. URL: https://doi.org/10.1093/molbev/msaa204 (cit. on pp. 19, 22, 26, 31).

[8] Rob Phillips, Jane Kondev, and Julie Theriot. *Physical Biology of the Cell*. New York: Garland Science, Taylor & Francis Group, Nov. 2008. ISBN: 978-0815341635. URL: http://www.worldcat.org/search?qt=worldcat_org_all&q=0815341636 (cit. on p. 23).

[9] Alan F. Rubin, Hannah Gelman, Nathan Lucas, Sandra M. Bajjalieh, Anthony T. Papenfuss, Terence P. Speed, and Douglas M. Fowler. «A statistical framework for analyzing deep mutational scanning data». In: *Genome Biology* 18.1 (Aug. 2017), p. 150. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1272-5. URL: https://doi.org/10.1186/s13059-017-1272-5 (cit. on p. 26).

[10] Sébastien Boyer, Dipanwita Biswas, Ananda Kumar Soshee, Natale Scaramozzino, Clément Nizak, and Olivier Rivoire. «Hierarchy and extremes in selections from pools of randomized proteins». In: *Proceedings of the National Academy of Sciences* 113.13 (2016), pp. 3482–3487. ISSN: 0027-8424. DOI: 10.1073/pnas.1517813113. eprint: https://www.pnas.org/content/113/13/3482.full.pdf. URL: https://www.pnas.org/content/113/13/3482 (cit. on pp. 26, 28).

[11] Douglas M. Fowler, Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. «High-resolution mapping of protein sequence-function relationships». In: *Nature Methods* 7.9 (Sept. 2010), pp. 741–746. ISSN: 1548-7105. DOI: 10.1038/nmeth.1492. URL: https://doi.org/10.1038/nmeth.1492 (cit. on p. 26).

[12] Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. «Adaptation in protein fitness landscapes is facilitated by indirect paths». eng. In: *eLife* 5 (July 2016). e16965[PII], e16965. ISSN: 2050-084X. DOI: 10.7554/eLife.16965. URL: https://doi.org/10.7554/eLife.16965 (cit. on pp. 26, 30).

[13] Jakub Otwinowski. «Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function». In: *Molecular Biology and Evolution* 35.10 (Aug. 2018), pp. 2345–2354. ISSN: 0737-4038. DOI: 10.1093/molbev/msy141. eprint: https://academic.oup.com/mbe/article-pdf/35/10/2345/27171842/msy141.pdf. URL: https://doi.org/10.1093/molbev/msy141.

[14] Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. «Machine learning-assisted directed protein evolution with combinatorial libraries». In: *Proceedings of the National Academy of Sciences* 116.18 (2019), pp. 8852–8858. ISSN: 0027-8424. DOI: 10.1073/pnas.1901979116. eprint: https://www.pnas.org/content/116/18/8852.full.pdf. URL: https://www.pnas.org/content/116/18/8852.

[15] 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.

[16] Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. «Inferring the shape of global epistasis». In: *Proceedings of the National Academy of Sciences* 115.32 (2018), E7550–E7558. ISSN: 0027-8424. DOI: 10.1073/pnas.1804015115. eprint: https://www.pnas.org/content/115/32/E7550.full.pdf. URL: https://www.pnas.org/content/115/32/E7550.

[17] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. «Deep generative models of genetic variation capture the effects of mutations». In: *Nature Methods* 15.10 (Oct. 2018), pp. 816–822. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0138-4. URL: https://doi.org/10.1038/s41592-018-0138-4.

[18] 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.