Politecnico di Torino

Master Degree in
Physics of Complex Systems

Using Glauber-Pseudolikelihood Dynamics to Generate Artificial Proteins

Politecnico
di Torino

Supervisors

Andrea PAGNANI

Anna Paola MUNTONI

Author

Marco CIPOLLINI

Academic year 2020/2021

Contents

Abstract

Since the fundamental processes governing the functioning of the biological cell were first discovered, proteins' behaviour became a subject of notable interest among the scientific community. The reason behind the relevance of this topic is justified by the fact that proteins are responsible for a numerous ensemble of different activities which are essential to the survival of all life forms present on this planet. Given this fact, it goes without saying that one of the first issues that biologists and scientists had to deal with was to analyze proteins in order to identify their functions inside the organism.

Proteins are macro-molecules constituted by organic monomers, the amino acids, linked in a chain that eventually, after its assembly in the cell, proceeds to fold into a particular three-dimensional structure, known as the tertiary structure of proteins. One of the most remarkable discoveries on the subject is that it is indeed this configuration, conveyed by inter-molecular forces between amino acids, that is responsible for the specific role of proteins inside the organism.

After examining several proteins by the means of x-ray spectroscopy [24], discerning their amino acidic sequence and structure and detecting their functionality, it was learned by the scientific community that many proteins obtained from even very different types of organism share a similar three-dimensional pattern. The most endorsed hypothesis is that these proteins have a common evolutionary origin, and while the mutation mechanisms at the basis of evolution gave rise in the millions of years since the emergence of life on this planet to the observable amino acids' variability seen in these sequences, the natural selection instead preserved their structure almost unaltered. It was then straightforward to collect all these proteins, called homologous, inside different families according to their function.

In order to model the tertiary structure of proteins, a series of statistical mechanics' techniques have been employed under the name of direct coupling analysis (DCA) [7], which allowed scientists to determine for each protein family a distinguishing set of parameters, by means of which it is possible to outline the folded configuration of amino acids' sequences.

While the problem of classifying proteins inside evolutionary-related families has been successfully solved since the 1970' [12], in more recent years a more demanding challenge has arised: that is to biosynthesize artificial sequences capable of mimicking a desired function associated with a certain protein family, and at the same time being more stable and efficient than the natural ones.

On the basis of the works of E.Aurell et al. [6],[19] and S.Cocco et al. [23], this paper aims at showing that it is possible to design a generative algorithm capable of building up original amino acids' sequences using parameters inferred from various protein domain families, with the purpose of reproducing the distinguishing three-dimensional structure and therefore the function of the family analyzed if ultimately assembled in laboratory.

An illustration of the way it is possible to infer these parameters from the protein families and how they serve the purpose of generating these sequences will follow in the upcoming chapters of this work.

1 Introduction

1.1 Aim of the thesis

The main objective of this thesis work was to develop a generative model capable of producing original amino acids' sequences that could replicate the main features and functions of the natural sequences if ultimately assembled in laboratory.
To be more specific, the main characteristic of the latters that these artificial proteins should possess is the same allocation of the direct contacts between couples of sites in the chain, which translates into the same three-dimensional structure. The importance of designing artificial proteins is emphasised by a series of applications that are gaining success in recent years, such as the development of a therapy against SARS-CoV-2 infections [14].
In order to accomplish the similitude with the natural proteins' structure, it was necessary to infer the parameters of my model from the natural proteins through the means of pseudolikelihood maximization [6] and use such parameters to build our sequences. To check the significance of my results and the performance of the generating algorithm, I decided to implement different statistical tools for this purpose, of which the main are described in section 2.2 of chapter 2.
All the results of my calculations are reported and commented in chapter 3, while for a conclusive observation see chapter 4.
As a final introductory remark, since all the functions used to infer the parameters where already translated into Julia language by prof. A. Pagnani prior to the beginning of my work, I decided to write all my code, which is available in its entirety at the web page `https://github.com/marcocippo97/Glauber-Pseudolikelihood-Dynamics.git`, in the same language.

1.2 Generative Models

In the framework of statistical classification, a generative model is a model for the conditional probability of observing variable X given a training dataset knowledge Y, i.e. $P(X|Y)$ [21].
In our picture, the observable X is representing the artificial sequence we aim at producing, the Y data refers to the collection of natural proteins in our possess, and the probability distribution $P(X|Y)$ describes the "fairness" of observable X, which we can interpret as the probability that our generated sequence is able to mimic the characteristic behaviour of the family of proteins Y.
This model is opposed to the discriminative (or conditional) one, which instead seeks to estimate the probability of assigning a certain label (i.e.classify) Y to an unobserved data X, namely computing $P(Y|X)$. Translated into our context, this would mean to assert in which protein family a certain amino acids' sequence belongs to.
Exploiting Bayes' rule, we immediately see that there is a direct relation between the two approaches, since

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{1}$$

where $P(X)$ and $P(Y)$ symbolize the marginal distributions of observable X and label Y, while $P(X, Y)$ is the joint probability of the two. Indeed the generation of samples, which in reality is related to the latter probability distribution, can be ideally accomplished by both type of models, but since the prior information $P(X)$ is generally hard to get, we resort to the generative model for this task and reserve the discriminative one for classification.

1.3 Proteins

Proteins are organic polymers constituted by a combination of twenty different monomers, the amino acids, which they bind into a linear chain of variable length, going from dozens to hundreds of thousands of constituents [1].
They are present in both prokaryotic and eukaryotic life forms and their different functions and purposes inside the organism are innumerable. To name a few, they catalyze as enzymes many vital processes in the cell, such as the cell polarization [20], or act as vehicles for intra-cell transport, or even grant mechanical stability to tissues.
They are assembled by ribosomal ribonucleic acids' (rRNA) and proteins' aggregates know as ribosomes, which, starting from a messenger RNA (mRNA) carrying the genetic information acquired from the DNA, catalyze the peptidic bonds that link the amino acids to each other.
Proteins are characterized by four different type of structures [4]: the first one is represented by their mere sequence of amino acids and the second one is constituted by local patterns that the amino acids may assume as a first consequence of inter-molecular forces. There are two types of patterns observable, that may also be present both in the same protein, which are the α-helix and the β-sheet, as they are shown in figure 1.

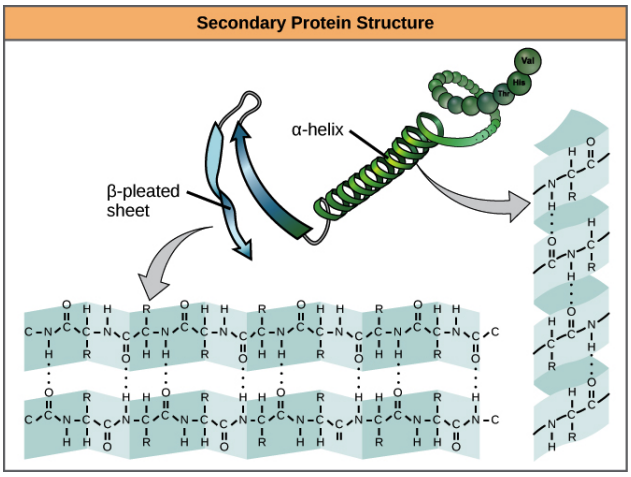

Figure 1: α-helices and β-sheets secondary structures in proteins

The third structure is the global three-dimensional shape that the protein ac-

quires after the folding process is finished, and the forth one occurs only in proteins' aggregates constituted by different amino acids' sequences linked together by weak forces or possibly by disulfide bridges, as happens for the hemoglobin protein, which is formed by four different components.
Of all types of protein structures we will be more interested in the third one, since it is responsible for how the protein interacts with its surrounding, determining its function in the cell. For example a receptor protein on the outside of the cell's membrane of a lymphocyte, due to its distinguishing structure, is capable of establishing a bond with only a particular target molecule, which sends a signal to the cell, helping it eventually to locate an antigen in the organism [13].
The three-dimensional conformation of the protein begins to form as early as the amino acids are linked together by the ribosome, and it is produced by an equilibrium process which is lead by the minimization of the Helmholtz free energy [11] of the protein, as it is known in thermodynamics, and involves a range of different possible inter-molecular interactions, going from the Hydrogen bond to the Van der Waals forces, that may engage amino acids far away from each other in the sequence.

1.4 Domain families

A notable characteristic of many proteins is that they simultaneously carry out several tasks inside the organism. This multi-tasking feature is due to a particular division of the amino acids' chain into sub-sequences which fold independently the ones from the others, forming the so called domains [2]. An example is provided in figure 2, where it is shown a depiction of a protein visibly made up of three distinct domains.
Being in general significantly different in shape from the other sub-units, each domain plays its role inside the cell without interfering with the others.
Since the first protein sequences were determined by Frederick Sanger over half a century ago [8], and consequently their conformations and tasks inside the cell were identified, it became clear that domains that fulfill a similar function in even widely different types of organisms share the same kind of tertiary structure. It then came naturally for the scientific community to start a project of collection and subdivision of protein domains into so called families, based on this feature, culminating in 1995 with the foundation of Pfam by Erik Sonhammer, Sean Eddy and Richard Durbin [12], a web site which contains so far 19,179 different families of protein domains, with new entries inserted every year.
To account for this phenomenon of similarity among proteins, we can consider the fact that all the living beings appeared on Earth had a common ancestor from whence they derived [22]. Through the evolutionary mechanisms that allow for environment's adaptation, each organism has speciated, while maintaining the structure of some macro-molecules (including proteins), responsible for basic and vital processes, almost unaltered. All genes, and the relative proteins they are coding for, that have a common ancestry are called homologous. To be more specific: if two different genes diverged from a duplication event of the cell, they are classified as paralogues, whereas if they diverged after a speciation event, which led to two different species of organisms, they are called orthologues.
To estimate the times these proteins appeared in their organisms and to better

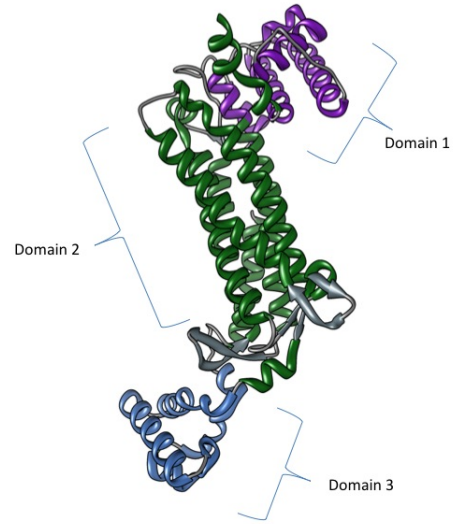

Figure 2: Domains division of a protein

clarify their ancestry relation, which is known as phylogeny, biologists began to use mathematical models to generate graphs, by the name of phylogenetic trees (e.g. figure 3), which helped to reconstruct their connections.

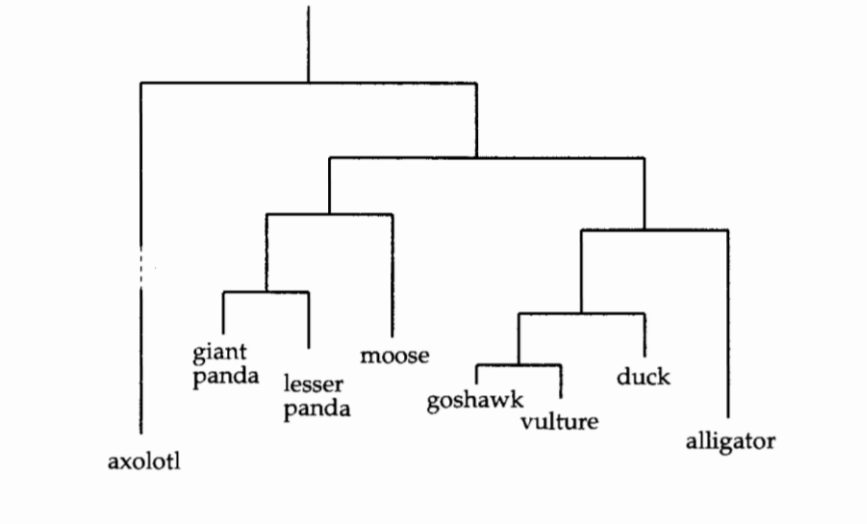

Figure 3: Phylogenetic tree linking different species of animals (Durbin [22])

Most of the models necessary for the phylogenetic tree's construction demand to determine the Levenshtein distance between two sequences of amino acids, which we will also refer to as residues from now on, consisting in the minimum number of single-character changes, to be chosen from insertions, deletions or substitutions operations, required to change one protein into the other.
With every "path", i.e. the succession of different edits relating two sequences, it is possible to associate a score, which we'll aim to maximize. At the end of such operation we will align the two sequences one on the top of the other, introducing a twenty-first character to our alphabet of residues, which is the gap sign "-", accounting for the deletion and insertion actions.
Applying this concept to a probabilistic perspective, meaning questioning ourselves what was the most probable sequence of edits which generated the alignment with maximum score, we can introduce a probabilistic model, known as Hidden Markov Model (HMM), to fulfill this task.
If we aim to collect a group of homologous proteins, we can apply a variation of this procedure, known as profile Hidden Markov Models (pHMM) [22], to generate a multiple sequence alignment (MSA), corresponding to a table in which the sequences are displayed horizontally one on the top of the other, as we can see in figure 4. Indeed all databases of family of proteins, such as the Pfam, regroup them as MSAs encoded in files with FASTA format [18].

```
structure:  ...aaaaa...bbbbbbbbbb.....ccccccCCC..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........+........++
```

Figure 4: Multiple sequence alignment of ten immunoglobins superfamily domains (Durbin [22])

2 Models and algorithms

2.1 Inference of the parameters

As we have seen in the introductory section, the tertiary structure of proteins is of notable interest to biologists, since it is the feature that provides the protein its functionality in the cell.
If we are interested in comparing sequences and subdivide them into families, we will have to associate a measurable quantity to each structure. The observable that will come handy in this situation is the couplings'/contacts' distribution, listing all the couples of sites in the chain which we'll consider as bounded by inter-molecular interactions. A frequently used criterion to discern bonds is a proximity, measured by x-ray spectroscopy, of $\sim$ 7 Angstrom, excluding the residues which are nearest neighbors along the sequence. Note that, due to the limited mobility of intra-molecular bonds inside each amino acid, there is a one-to-one correspondence between a list of contacts and the three-dimensional shape of our protein.
Meanwhile, if we want to detect the bonds between amino acids starting from a multiple sequence alignment, a series of methods have been devised, which gather under the name of direct coupling analysis (DCA) and employ the tools of physics and statistics, especially the concept of correlation, to address this issue [7]. The reason why the term "direct" is included in DCA is related to the fact that site-to-site correlations could arise in principle through a "real" (direct) contact between the two amino acids in the chain or by a network-mediated indirect interaction, which could lead to an erroneous interpretation of the structure of the protein. The goal of this methods is to discern between these two types of contacts.
To better visualize the difference between these contacts, we can look at figure 5: we can see from this depiction of a folded filament of protein that there are contacts between amino acids A and B, and between amino acids C and D (left image). If the protein is subjected to an external stimulus, e.g. heat or electromagnetic waves, the bond between C-D could break (right image), causing a separation, because of the chain conformation, between also residues A and B. Analyzing the data associated to these two situations, we would certainly measure a correlation between A-B residues and between C-D residues, which are the two direct contacts inside the structure, while we could also measure, because of the previous reasoning, a correlation between A-C or A-D, which are instead indirect contacts, since there is no real physical proximity between these residues.
We already know that domain families group together different proteins, yet similar in terms of contacts' distribution. If we want to apply a probabilistic model in order to justify this variety of sequences, we will have to consider a MSA as a sample extracted from a Boltzmann probability distribution (with β assumed equal to 1)

$$P(\bar{\sigma}|\mathbf{J},\mathbf{h}) = \frac{1}{Z} exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1\leq i<j\leq N} J_{ij}(\sigma_i,\sigma_j)\right) \tag{2}$$

where the vector $\bar{\sigma}$ represents our sequence of amino acids, each one taking scalar values from 1 to 21 (we recall that the twenty-first number refers to the

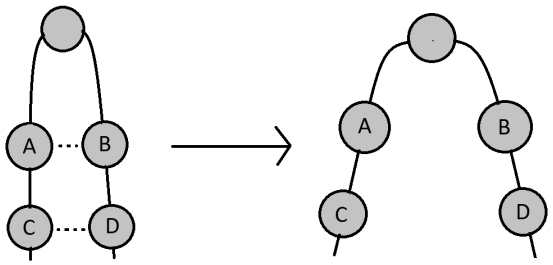

Figure 5: A representation of direct and indirect contacts among residues

gap inside the MSA), and $\{h_i(\sigma_i),\ J_{ij}(\sigma_i, \sigma_j)\}$ will be the parameters of our exponential function.
The "field" parameter $h_i(\sigma_i)$ is responsible for highlighting a possible conservation of a residue in a particular position of our chain, which translates to a high empirical frequency of the corresponding letter in the column of the MSA, while the "coupling" parameter $J_{ij}(\sigma_i, \sigma_j)$ takes into account the direct or indirect contacts between residues. Indeed if two amino acids are bounded together by inter-molecular forces and one of them is brought to a change by a mutation, it makes sense that the other one will change too, establishing a correlation between the two sites.
The statistical physics model introduced with the Hamiltonian

$$H(\bar{\sigma}) = \sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leq i < j \leq N} J_{ij}(\sigma_i, \sigma_j) \tag{3}$$

is called the Potts model [25], a generalization of the Ising model with q=21 possible "spin" states. To validate the choice of this particular distribution, we could resort to the maximum entropy principle.
Assuming the fact that our unknown distribution $P(\bar{\sigma})$ must fulfill some constraints, we aim at deriving the distribution which respects the latters, while not introducing any other bias, which translates to the maximization of the function

$$\begin{aligned} \Gamma(P) &= S(P) \\ &+ \sum_{i<j} \sum_{a,b} \lambda_{ij}(a,b) \left(P_{ij}(a,b) - f_{ij}(a,b)\right) \\ &+ \sum_{i} \sum_{a} \lambda_{i}(a) \left(P_{i}(a) - f_{i}(a)\right) \\ &+ \mu \left(1 - \sum_{\bar{\sigma}} P(\bar{\sigma})\right) \end{aligned} \tag{4}$$

where the function $S(P)$ is the Shannon entropy, measuring how much our distribution P differs from a uniform distribution (it has indeed the form of a

Kullback-Leibler divergence between the latter and $P(\bar{\sigma})$, except for a change of sign and a constant additive term, see Appendix A for more details), and corresponds to

$$S(P) = -\sum_{\bar{\sigma}} P(\bar{\sigma}) log(P(\bar{\sigma})) \tag{5}$$

The parameters $\{\lambda_{ij}, \lambda_i, \mu\}$ are instead the Lagrange multipliers associated to the constraints

$$\begin{cases} \sum_{\bar{\sigma}} P(\bar{\sigma}) = 1 \\ P_{ij}(a,b) = f_{ij}(a,b) \\ P_i(a) = f_i(a) \end{cases} \tag{6}$$

The first one is non other than the usual normalization condition of probability distributions, while the other two impose the equality between the one-point and two-point empirical frequencies of the MSA and the marginal distributions of $P(\bar{\sigma})$

$$\begin{cases} P_i(a) = \sum\limits_{\bar{\sigma}, \sigma_i = a} P(\bar{\sigma}) \\ P_{ij}(a,b) = \sum\limits_{\bar{\sigma}, \sigma_i = a, \sigma_j = b} P(\bar{\sigma}) \\ f_i(a) = \frac{1}{M} \sum\limits_{m=1}^{M} \delta(\sigma_i^{(m)}, a) \\ f_{ij}(a,b) = \frac{1}{M} \sum\limits_{m=1}^{M} \delta(\sigma_i^{(m)}, a) \delta(\sigma_j^{(m)}, b) \end{cases} \tag{7}$$

where M is the number of sequences in our MSA, and $\delta(x,y)$ is the Kronecker delta function. If we now impose the minimization condition $\frac{\partial \Gamma(P(\bar{\sigma}))}{\partial P(\bar{\sigma})} = 0$ we would get the same Boltzmann distribution as in equation (2), with the change of variables

$$\begin{cases} J_{ij}(a,b) \leftarrow \lambda_{ij}(a,b) \\ h_i(a) \leftarrow \lambda_i(a) \end{cases} \tag{8}$$

The branch of study developed to solve the problem of determining the set of parameters $\{\mathbf{J}, \mathbf{h}\}$ that suits the most the homologous sequences takes the name of inverse statistical physics [23]. The "inverse" term refers to the opposite flow of information that occurs with respect to classical statistical mechanics, in which indeed from known parameters we are able to derive the desired observable.
In the framework of Bayesian probability, in order to fulfil the previously mentioned task, we have to maximize the likelihood associated with the proteins, which corresponds to the conditional probability of sampling this exact set of

sequences given the parameters. More precisely we have to determine

$$\{\mathbf{J}^*, \mathbf{h}^*\} = \arg\max_{\mathbf{J},\mathbf{h}} P(\boldsymbol{\sigma}|\mathbf{J}, \mathbf{h}) \tag{9}$$

where the probability $P(\boldsymbol{\sigma}|\mathbf{J}, \mathbf{h})$ is referring to our whole MSA

$$\boldsymbol{\sigma} = \{\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(M)}\}$$

and each $\sigma^{(\cdot)}$ denotes a different natural sequence.
Since we expect, given the huge number of possible datasets and the normalization constraint of the distribution, that the probability value could turn out particularly low, causing problems of underflow in our computers, we replace the likelihood with a rescaled negative log-likelihood (exploiting the factorization of the former)

$$l = -\frac{1}{M}\sum_{m=1}^{M} log(P(\sigma^{(m)}|\mathbf{J}, \mathbf{h})) \tag{10}$$

This operation is legitimate because the logarithm is a monotonic function and it will not affect the search of extrema points, while introducing a minus will just switch the argmax into an argmin.
In order to determine the log-likelihood in equation (10), we will have to apply the constraints (6), with a slight modification of the empirical frequencies functions

$$\begin{cases} f_i(a) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(a, \sigma_i^{(m)}) \\ \\ f_{ij}(a,b) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(a, \sigma_i^{(m)}) \delta(b, \sigma_j^{(m)}) \end{cases} \tag{11}$$

The introduction of the constants $\{\omega_m\}$ and M_{eff} is due to the fact that we can't consider our samples to have been independently drawn one after the other from the distribution $P(\bar{\sigma})$, since evolutionary times separating sequences in the phylogenetic trees could be too short to grant uncorrelation between two following proteins. Indeed we will assign a weight ω_m to each sequence to take into account this correlation, and renormalize the empirical frequencies by the sum of all weights $M_{eff} = \sum_{m=1}^{M} \omega_m$. The exact way this weight assignment is done goes beyond the purpose of this preliminary part; all that is needed to know is that we will assign smaller weights to sequences which share a similarity with a good number of other homologous proteins, while the more original and unique sequences will have greater weights.
At this point, due to the imposition of the set of equations regarding the empirical frequencies, we realize that there is an overparametrization of our problem, since a number of different sets of parameters $\{\mathbf{J}, \mathbf{h}\}$ will lead to the same probability distribution $P(\bar{\sigma})$. In fact, among all $Nq + \frac{N(N-1)}{2}q^2$ possible parameters, it can be demonstrated that only a subset of $N(q-1) + \frac{N(N-1)}{2}(q-1)^2$ nonredundant ones is needed. To fix this overparametrization we can apply a gauge choice that leaves the distribution unaltered, imposing

$$\sum_{s=1}^{q} J_{ij}(s,k) = \sum_{s=1}^{q} J_{ij}(k,s) = \sum_{s=1}^{q} h_i(s) = 0 \ \forall \ i,j,k \tag{12}$$

known as the Ising, or zero-sum, gauge.
Having dealt with this issue we have finally a differentiable form for our negative log-likelihood

$$l = log(Z) - \sum_{i=1}^{N}\sum_{k=1}^{q} f_i(k)h_i(k) - \sum_{1\leq i<j\leq N}\sum_{k,l=1}^{q} f_{ij}(k,l)J_{ij}(k,l) \tag{13}$$

from which, taking its partial derivatives with respect to the parameters and imposing them to be equal to 0, in principle it is possible to trace its minima points.
The real struggle with this model comes with the calculation of the partition function Z, which the enormous number of variables into play makes it impossible to accomplish with standard computational methods.
To bypass this operation E.Aurell et al. proposed in their article [6] to substitute the usual likelihood with a pseudo-likelihood describing the probability for a given residue to appear in a certain position along the chain, conditional to the fact of having observed all of the other residues, which takes the form

$$P(\sigma_r^{(b)}|\sigma_{-r}^{(b)}) = \frac{exp\left(h_r(\sigma_r^{(b)}) + \sum\limits_{i\neq r,i=1}^{N} J_{ri}(\sigma_r^{(b)},\sigma_i^{(b)})\right)}{\sum\limits_{l=1}^{q} exp\left(h_r(l) + \sum\limits_{i\neq r,i=1}^{N} J_{ri}(l,\sigma_i^{(b)})\right)} \tag{14}$$

We immediately notice from this formula that the complicated partition function disappears in deriving this marginal distribution, substituted by a more manageable one.
An inconvenience with this model is that this probability measure relies only on the parameters $\mathbf{J_r} = \{\mathbf{J_{ri}}\}_{i\neq r}$ and $\mathbf{h_r}$ associated with node r. Applying the minimization of the negative log-pseudolikelihood

$$g_r(\mathbf{J_r},\mathbf{h_r}) = -\frac{1}{M_{eff}}\sum_{m=1}^{M} \omega_m log(P(\sigma_r = \sigma_r^{(m)}|\sigma_{-r} = \sigma_{-r}^{(m)})) \tag{15}$$

then will in general lead to different predictions for parameters $\mathbf{J}_{rq}$ and $\mathbf{J}_{qr}$, depending on which site we minimize over. To fix this problem we can resort to an average of the two $\frac{\mathbf{J}_{rq}+\mathbf{J}_{qr}}{2}$ or we can minimize over the sum of all site-dependent pseudolikelihoods

$$l_{pseudo} = -\frac{1}{M_{eff}}\sum_{r=1}^{N}\sum_{m=1}^{M} \omega_m log(P(\sigma_r = \sigma_r^{(m)}|\sigma_{-r} = \sigma_{-r}^{(m)})) \tag{16}$$

where we can see that in the formulas above we added the weight factors ω_m and M_{eff} introduced previously.
If we consider an ordinary protein family MSA with an alignment length of the

order of 10^2 residues, we can determine that the number of necessary parameters to be trained over this dataset is of the order of 10^6. Since the number of proteins included in a family is usually in a far modest range, we have to account for the possibility of an overfitting. To prevent this eventuality, we may apply a regularization R_{l_2} over our log-pseudolikelihood, imposing

$$\{\mathbf{J}^*, \mathbf{h}^*\} = \arg\max_{\mathbf{J},\mathbf{h}} \{l_{pseudo}(\mathbf{J}, \mathbf{h}) + R_{l_2}(\mathbf{J}, \mathbf{h})\} \tag{17}$$

where the regularization term takes the form

$$R_{l_2}(\mathbf{J}, \mathbf{h}) = \lambda_h \sum_{i=1}^{N} ||\mathbf{h_i}||_2^2 + \lambda_J \sum_{1 \le i < j \le N} ||\mathbf{J_{ij}}||_2^2 \tag{18}$$

The symbol $||...||_2$ refers to the Frobenius norm, while the two constants λ_h and λ_J are the regularization parameters to be tuned, possibly even independently one from the other, in order to get the desired result from our calculation. In my specific case, both the parameters were set to the value $\lambda_h = \lambda_J = 0.01$, since it is a widely used choice in this context and proved to work well for my purpose.
Now that we have defined a method through which is possible to determine the field and couplings parameters that minimize the negative log-pseudolikelihood, we have to deal with the problem of assigning a score to each couple of sites in the alignment. This operation is required, within the framework of DCA, if we want to discern the direct contacts inside the tertiary structure of the protein domain and it will be vital, when we'll discuss the sequence generating algorithm, to verify the accuracy of the latter, since it represents our estimation on how probable each couple of sites in the chain is indeed a real contact in the three-dimensional configuration.
First of all we notice that a scoring function should assign to each couple of sites in the chain a scalar, while the coupling's parameter $\mathbf{J_{ij}}$ is on the other hand a tensor. It was then proposed again by E.Aurell in article [6] to implement this scoring function

$$S_{ij} = S_{ij}^{FN} - \frac{S_{\cdot j}^{FN} S_{i\cdot}^{FN}}{S_{\cdot\cdot}^{FN}} \tag{19}$$

where the dots substituting a letter indicate an average taken over that position and S_{ij}^{FN} is a different estimation of the score based on the Frobenius norm

$$S_{ij}^{FN} = ||\mathbf{J'_{ij}}||_2 = \sqrt{\sum_{k,l=1}^{q} J'_{ij}(k,l)^2} \tag{20}$$

Since the regularization fix the parameters to a certain gauge, we have replaced the usual coupling matrix $J_{ij}(k,l)$ with a corrected form

$$J'_{ij}(k,l) = J_{ij}(k,l) - J_{ij}(\cdot,l) - J_{ij}(k,\cdot) + J_{ij}(\cdot,\cdot) \tag{21}$$

to restore the wished Ising gauge.

As a final remark of this section we are now able to say that it is possible through the means of pseudolikelihood maximization to gain information about the folded structure of a protein family, and all the code implementing the functions necessary for the purpose is available in Julia language at the web page `https://github.com/pagnani/DynamicPLM.jl.git` with kind permission of prof. A. Pagnani.

2.2 Observables of interest

In order to understand if the generative algorithm samples artificial protein sequences which are related, and so in principle could have a similar behaviour if ultimately assembled in laboratory, to the natural ones belonging to the family we are considering, we can estimate a few quantities of interest extractable from the data.

2.2.1 Connected correlation

To see if there is a similar correlation, in terms of residues' values, between positions in the chain in both the natural and the sampled sequences, we can compute the two-point $c_{ij}(a,b)$ and three-point $c_{ijk}(a,b,c)$ connected correlations.
It is important to notice that taking into consideration the three-point connected correlation is a particularly demanding test for our algorithm, since in the Potts model there are no parameters to determine which reflect an interaction among three residues in the chain.
We start off by defining the single site, two site and three site empirical frequencies

$$\begin{cases} f_i(a) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(\sigma_i^{(m)}, a) \\ \\ f_{i,j}(a,b) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(\sigma_i^{(m)}, a)\delta(\sigma_j^{(m)}, b) \\ \\ f_{i,j,k}(a,b,c) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(\sigma_i^{(m)}, a)\delta(\sigma_j^{(m)}, b)\delta(\sigma_k^{(m)}, c) \end{cases} \tag{22}$$

Then at this point we can evaluate for both the natural and sampled sequences the previously mentioned connected correlations

$$c_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b) \tag{23}$$

$$\begin{aligned} c_{ijk}(a,b,c) = {} & f_{ijk}(a,b,c) - f_{ij}(a,b)f_k(c) \\ & - f_{ik}(a,c)f_j(b) - f_{jk}(b,c)f_i(a) \\ & + 2f_i(a)f_j(b)f_k(c) \end{aligned} \tag{24}$$

and evaluate the Pearson correlation coefficient $\rho(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ (where σ_x denotes the variance of x and σ_{xy} denotes the covariance of x and y) between all the pairs of connected correlations of sampled and natural sequences associated

with the same values of position indices (i, j, k) and residues' values (a, b, c).
If the algorithm generates sampled sequences with similar connected correlations to those of the natural ones, then the Pearson correlation coefficient will be close to one.
In order to compare graphically the similarity between the two correlations, we'll set up a plot showing on the x axis the natural sequences' correlations of all the couples (or triplets, if it is the case of the three-point one) of sites/residues considered, then for the y axis we'll apply the same procedure to the sampled sequences, respecting the same order of couples/triplets taken previously. Finally, we draw on the plot the identity function for reasons of comparison. The more the two correlations are alike, the more the scatter plot will overlap the identity function.
As an example, we can look at figure 6, where two vectors, x and y, are compared. The two were generated in the same way, adding to each of the integers between 1 and 100 a random positive number up to its tenth percent. As we should expect the greater the number, the wider is the distance between the dots and the straight line.

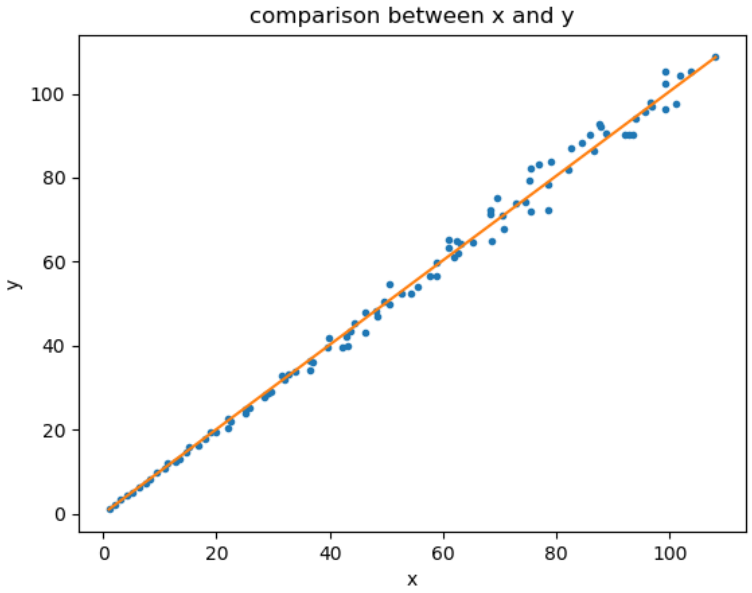

Figure 6: An example of a comparison between two vectors

2.2.2 Positive predictive value curve

Once the generative model has produced a sufficient number of artificial proteins, we can re-perform the maximum pseudolikelihood process onto this set of sampled sequences in order to obtain a new set of parameters $\{\mathbf{J_s}, \mathbf{h_s}\}$, to which, along with the previously calculated natural sequences' ones, we'll apply

the scoring function (19).
After evaluating the scores of all possible couples of positions in the chain for both the sampled and natural sequences, it is possible to verify how well this score evaluation matches with the real 3D folded structure of natural sequences. Using the `compute_roc` function from `DynamicPLM` package, it is now possible to determine the rate of correctly predicted direct contacts in the chain by relating the scores of each couple of residues to a known and previously determined (via x-ray spectroscopy) dataset, describing the typical structure of the domain family considered. This rate, known as "precision", will represents in our graph the fraction of correctly guessed contacts up to that point in the process, as it will be soon clarified.
To be more specific: the domain family dataset will contain a matrix associating each couple of sites with its relative distance in Armstrong, and our scores will be set in decreasing order. The algorithm proceeds by considering each score, starting from the highest one, and verifying if the couple of sites to which this evaluation belongs is a real contact in the structure, therefore if its distance in the dataset is at most 7 Angstrom. We keep in mind that, regardless of their proximity in terms of Angstrom, couples of sites distant less than 5 positions along the chain are not considered for the precision evaluation, since due to protein folding, a non-trivial direct contact is highly unlikely.
Plotting this quantity with the number of each couple on the x axis and the respective precision (labelled on the graph as P_{ij}) on the y axis, we get the so called positive predictive value (PPV) curve.
This operation can be applied to both the set of sampled sequences and the set of natural ones, and by comparing the shapes of the two curves (and particularly the number of perfect scores, which usually is expected to be greater for the latter set for obvious reasons) we can estimate how good our generating function behaves.
In all the PPV curves that will be shown in the chapter dedicated to my results, I decided to focus only on the first 250 highest scoring couples, since the first deviation from a perfect prediction is well visible at this scale for both natural and generated sequences. However, in order to have a broader view at least once, we can see in figure 7 a full-scale PPV plot for the PF00014 protein family, showing the two curves related to the natural and sampled sequences and the perfect prediction one, consisting in a straight line (TP=1) until the number of contacts runs out (in this particular case was 375), then it decreases as $\frac{1}{x}$. As a final remark, note that the three curves ultimately overlap in the same point in the graph, which value corresponds to the probability of detecting a contact couple at random, i.e. by dividing the real number of contacts by the number of possible ones ($\binom{n}{2}$ with n equals to the alignment length, minus the number of couples closer than 5 locations).
By looking at this image, we have to bear in mind that it is more appropriate to estimate the performance of the generating algorithm by comparing the curve related to the sampled sequences with the naturals' one, instead of comparing it with the perfect prediction plot. Neglecting that the latter is referred to an ideal case, which most likely is not achievable in practice, the fact that the majority of direct contacts are not predicted by our algorithm up to a certain time in the process doesn't necessarily mean that it's not working well, since the outcome is inevitably bounded to the chosen definition of scoring function (shared by both the results concerning natural and generated sequences) and

the three-dimensional basic structure could be achieved anyhow with this limited set of contacts.
To highlight the well-functioning of the generative algorithm, I inserted in figure 7 also the straight line representing the performance of an algorithm that chooses each contact in the structure at random.

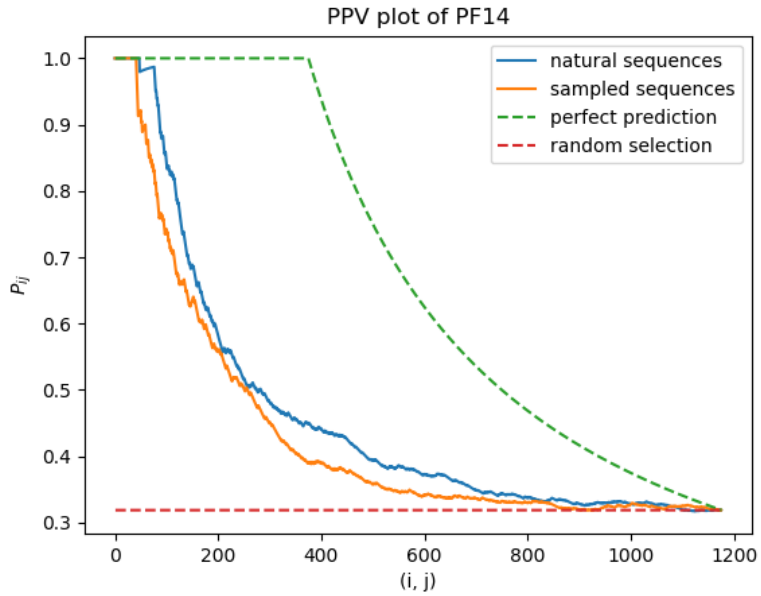

Figure 7: PPV plot for the PF00014 protein family

2.2.3 PCA plot

Another way to verify if the generated data behaves similarly to the natural sequences is to resort to the Principal Component Analysis (PCA) [16], a technique developed at the beginning of the twentieth century with the purpose of reducing the number of variables to deal with in a statistical problem, while maintaining only the most informative ones.
The method works in this way: we can estimate from the natural sequences (which can be represented as vectors in a n-dimensional Euclidean space, where n is the length of the protein chain) the two directions of maximal variance, first and second principal components (we can consider up to to n components in general), then we proceed by evaluating the projections of both natural and sampled sequences along these directions and at last we plot the projections' values into a 2D histogram.
After performing this calculations for the two types of data, by comparing the resulting histograms we can make an estimate about the goodness of the generating algorithm, since we expect the resulting images to be similar in both shape and color distribution.
In order to have a more adequate set of sampled sequences, I decided to ex-

clude repeated sequences from this calculation, since keeping them could alter the principal components' distribution, possibly leading it to a far different one with respect to the natural sequences' (which trivially are all different among them); so we have to take into account that the numbers of sequences in the two sets may not be of the same order.
As one last comment, we can look at figure 8 for an easy example of PCA application. The blue dots represent the original two-dimensional observations of a given problem, the direction u_1 is the direction of maximal variance for the variables, while the blue squares portray the projections of the latters along direction u_1, which are now one-dimensional as opposed to the original data.

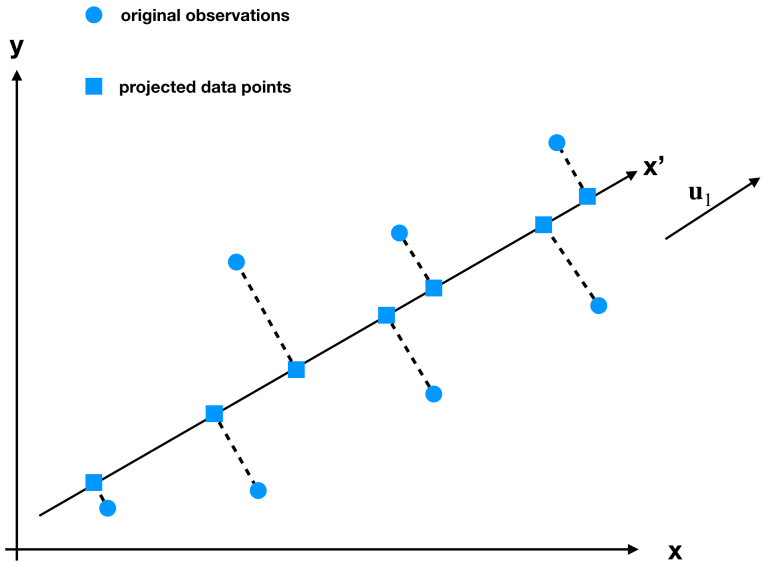

Figure 8: Example of PCA projections

2.3 Glauber-Pseudolikelihood Dynamics

As a starting point I based my generative model on the Glauber dynamics model [10] concerning the computational simulation of Ising systems in statistical physics. Indeed I decided to name this method Glauber-Pseudolikelihood Dynamics (GPD).
As the Glauber model, we are considering a discrete Markov chain [9] in which now the states are represented by the q^N possible configurations of our sequence of residues, where $q = 21$ is the number of amino acids (plus the gap) available and N is the length of the alignment of the MSA in which the natural sequences are collected.
We can make now the assumption that in long times our Markov chain dynamics will lead us to an equilibrium subset of our space of configurations, in which all the sequences share a similar 3D structure and the same functionality, whether they correspond to an already existing protein or an original one (in which we're

more interested). To make this hypothesis more reliable I decided to impose the transition probabilities of this network equal to the pseudolikelihood distribution already discussed in section 2.1.
In fact the coupling and field parameters $\{\mathbf{J}, \mathbf{h}\}$ showing in formula (14) will be the ones inferred from the homologous dataset through pseudolikelihood maximization, using prior the functions of the `DynamicPLM` package to carry out this calculation.
The edges of the network to which these probabilities are associated, link, as will be clarified shortly in the next section, every single sequence to the ones that differ by only a residue's value, including also a loop on the state itself. In order to have an idea on the type of structure we are building, we can look at figure 9 for a simplified version of our dynamics, in which the sequences have a length of three and our amino acidic alphabet will consists in only two letters, "A" and "B".

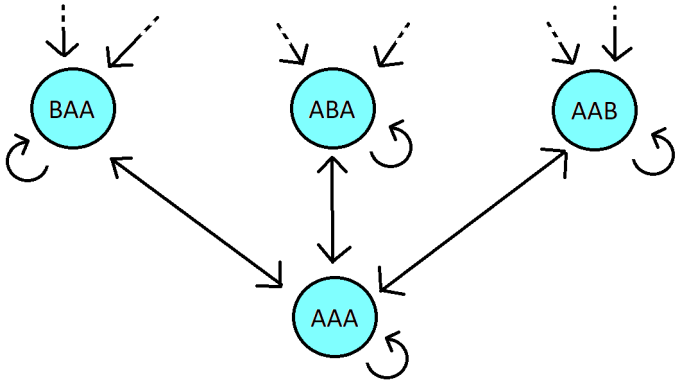

Figure 9: Simplified graph of the Markov chain dynamics

This is a fairly strong assumption, but given the fact that the parameters in formula (17) are intended to represent at best the peculiarities of the protein family, I still expected the generative model to work well enough.
It is important to specify that we do not have the presumption to describe with our Markov chain dynamics the evolutionary process from which the natural proteins physically derived, since we are only able to access the outer leaves of the phylogenetic tree describing the domain family, and not the inner nodes leading to the origin of the process, making every statement of this kind unverifiable.
The protein sequences were generated through two distinct algorithms, both to be considered part of the Monte Carlo class of sampling techniques [5]: the "asymptotic" and "contrastive" methods (a clarification about these names will follow up in the section concerning the latter method).

2.3.1 Asymptotic method

The asymptotic method is characterized by a relative long-time process, and can give the user an insight about how well the generating function of sequences performs after moving far away, in terms of amino acid substitutions, from its starting point in the space of configurations.
The first step of the algorithm is to generate a sequence with amino acids chosen uniformly at random, and of a length equal to the considered family's alignment length. Then, selecting a random site k at time t, a new sequence is generated at time $t+1$ changing the k's residue through the probability distribution

$$\sigma_k^{t+1} \sim P(\sigma_k|\sigma_{-k}^t, \mathbf{J}^{PLM}, \mathbf{h}^{PLM})$$

where the parameters $\{\mathbf{J}^{PLM}, \mathbf{h}^{PLM}\}$ are the ones derived through pseudolikelihood maximization (note that a residue remaining unchanged after a residue "flip" is allowed).
The sampling of the desired number of sequences, which is chosen to be of the same order of magnitude of the total number of natural sequences in the family, starts after a certain number of flips, or single residue changes, to be determined through a thermalization step. Then, after setting a sampling interval, always in terms of flips, in such a way that the generated sequences far by this quantity in time can be considered uncorrelated, the sampling of the sequences begins.
The thermalization step, whose name recalls the process in statistical physics through which a system reaches thermal equilibrium, is implemented similarly as the main method, but at each residue flip the value of the log-pseudolikelihood $l_{pseudo}(\bar{\sigma}, \mathbf{J}^{PLM}, \mathbf{h}^{PLM})$ of that configuration is saved in memory. After a sufficient number of flips the procedure ends, and the plot of the log-pseudolikelihood as function of time is displayed.
Looking at the graph is possible to notice at a certain time a saturation of l_{pseudo}, which is a good indicator for our sampling to begin, since we expect to be near the equilibrium state in our space of configurations.
The sampling interval is instead determined in a different way, involving the concept of Hamming distance between sequences.
After having reached the equilibrium in our space of configurations, a sufficient number of sequences is sampled continuously, that is without any waiting time between a sampling and the following one (as opposed to the main algorithm, as I will explain shortly). Then we can introduce the concept of autocorrelation

$$C(\Delta t) = \frac{1}{Tmeas - \Delta t} \sum_{t=1}^{Tmeas-\Delta t} d(t, \Delta t)$$

with $Tmeas$ being the number of sampled sequences and $d(t, \Delta t)$ being the Hamming distance function, counting the number of different residues at each position in the chain between the two sequences sampled at time t and at time $t + \Delta t$.
Also in this case plotting $C(\Delta t)$ as a function of Δt shows a saturation of the curve approximately at a point Δt^*, which can be taken as our sampling interval in the main algorithm, since samples generated at this frequency can be considered approximately uncorrelated.

2.3.2 Contrastive method

The contrastive method, differently from the asymptotic one, starts from a natural sequence belonging to the considered family of proteins and then applies the residue changing function for a relatively small number of times.
Note that by beginning our sampling from a natural sequence we are allowed to skip the thermalization step, which would be otherwise necessary for the asymptotic one, since we are already "close" to the equilibrium in our space of configurations.
As we will see later on in the Results paragraph all the quantities of interest calculated for this method are compared to the ones obtained through the asymptotic one. This last name hints to the fact the asymptotic method performs a much greater number of residue flips, such that it can be considered as an "extension" of the contrastive one, which instead takes its name from contrastive divergence learning [3].
Indeed the latter can give the user some information about how far in terms of Hamming distance the sampled sequences can get from the natural ones by performing such a restricted number of residue flips.
We expect the whole generating function to work well if such distance is not negligible, since in this case we could build up artificial proteins with a sufficiently original amino acid sequence and a site-to-site correlation similar to the one of the natural proteins we started off, meaning a good DCA prediction.
Going more into details, the algorithm starts from each natural sequence in the family and then applies the residue changing function once at each position in the chain. After having exhausted all the positions in the chain, the whole process is repeated a number of times equal to a low power of 2 (I chose to implement this method from the zeroth to the forth power). We will call "sweeps" these changes that involve each residue in the protein once. Then, the resulting sequence is sampled, giving eventually a sampled sequence for each natural one in the family.

3 Results

3.1 Protein family PF00014

To be more clear and to better report all the results of my calculations, I decided to introduce first the data related to the protein family PF00014 (which I will abbreviate as PF14 from now on to have a more compact format, especially in plots), which is the first dataset I analysed. I will proceed then by adding the three other families' results, and by making a comparison between these ones and the PF14's.
The information about all these protein families were picked as multiple sequences alignments from the Pfam website [12].
The PF14 protein family is characterized by a protein chain's length of $n = 53$ residues, and the number of natural proteins included in this family is 8871.
Starting off with the asymptotic method, two preliminary steps are needed before the sampling can actually begin. The first step to implement is the thermalization step, so as I already discussed we look for a saturation of the log-pseudolikelihood in our plot (the straight line indicates the last value of log-pseudolikelihood for the sake of comparison).
To generate a more easily understandable plot I performed a block division of the data, see Appendix B for details.
Looking at figure 10 we see by eye that a number of residue flips of the order of 10^6 is a reasonably good thermalization interval for the PF14 family.

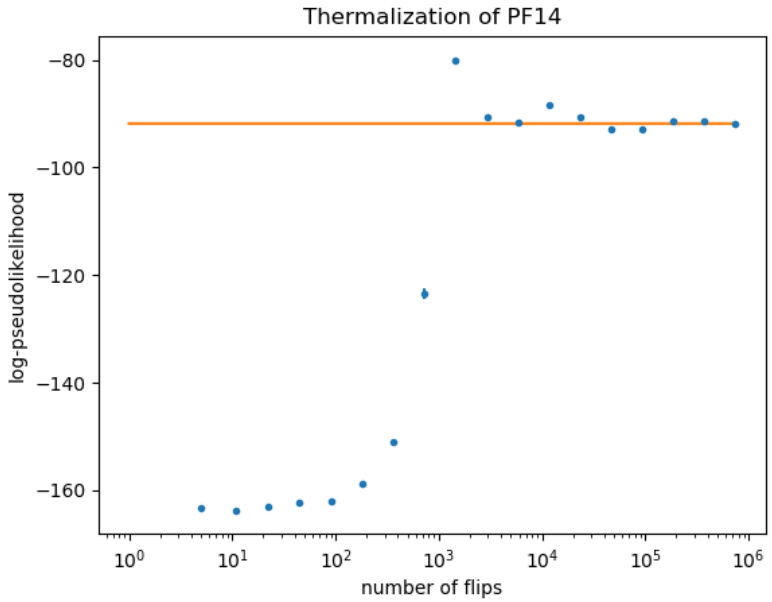

Figure 10: Thermalization curve of the PF14 protein family

The next step is to perform the autocorrelation analysis, and implementing the algorithm already discussed we see in figure 11 that the saturation of the curve happens approximately at $\sim 10^3$ flips, meaning that sequences sampled with

this number of residues' changes in between are to be considered uncorrelated. In figure 11 are also displayed as straight lines the last autocorrelation value (green line) and the mean (Hamming) distance among all natural proteins (orange line), as we expect the data to get close to this value as the number of residue flips increases.

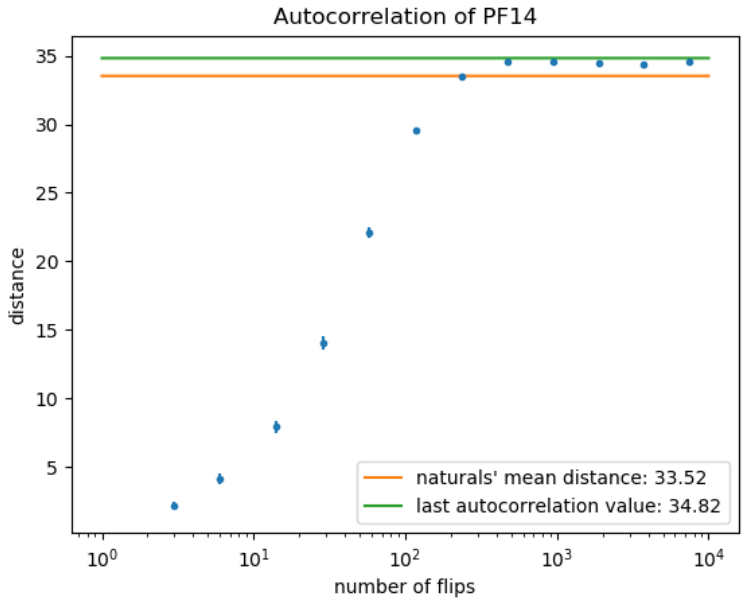

Figure 11: Autocorrelation curve of the PF14 protein family

Now that these two preliminary steps are done we can implement the main algorithm.
After having sampled a sufficient number of artificial sequences, we can determine the empirical two-point connected correlation for all the $\binom{n}{2}$ couples of indices in the chain.
In figure 12 it is possible to see a connected correlation comparison between the natural sequences' values and the generated ones, with their Pearson correlation coefficient being $\rho = 0.94$. The value of the latter and the fairly resemblance of the plot with the identity function suggest us that the second order statistics of the natural sequences is well reproduced in the generated sequences, but in order to validate the efficiency of the generating algorithm a further check by other means is necessary, e.g. by looking at the positive predictive value curve.

The PPV curve evaluation shows us in figure 13 a good amount of perfect scores ($P_{ij} = 1$) shared by both the generated and the natural sequences, with the latters having higher values of precision (as expected, being real sequences) for all the first 250 residue couples with higher scores analysed.
At last, we give a look at the principal component analysis (PCA) of the data and its histogram representation.
After performing the PCA evaluation for both the natural and sampled sequences, we can put side by side the two 2D histogram as in figure 14 to check

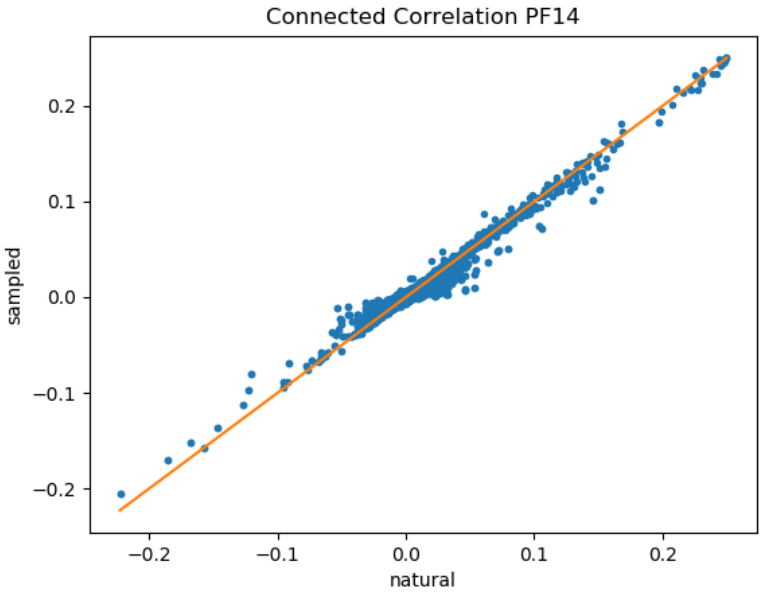

Figure 12: Connected correlation comparison of the PF14 protein family

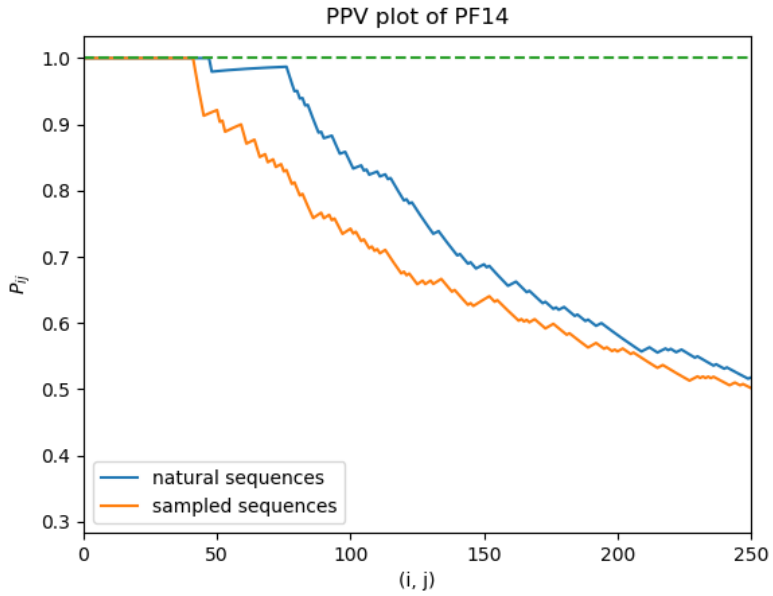

Figure 13: PPV curve of the PF14 protein family

the similarity between the two.
It can be easily seen that both the two plots' shape and color distribution

(representing the sequences number distribution) are alike, with the former differentiating for a well-marked border between the two "islands" of sites in the natural plot, and the latter having a broader light-colored zone for the sampled sequences' plot in the central area possibly due to the different number of sequences included in that set and the different colormap (visible on the right side) used to represent the data, besides a predictable partial inaccuracy of the sampling method.

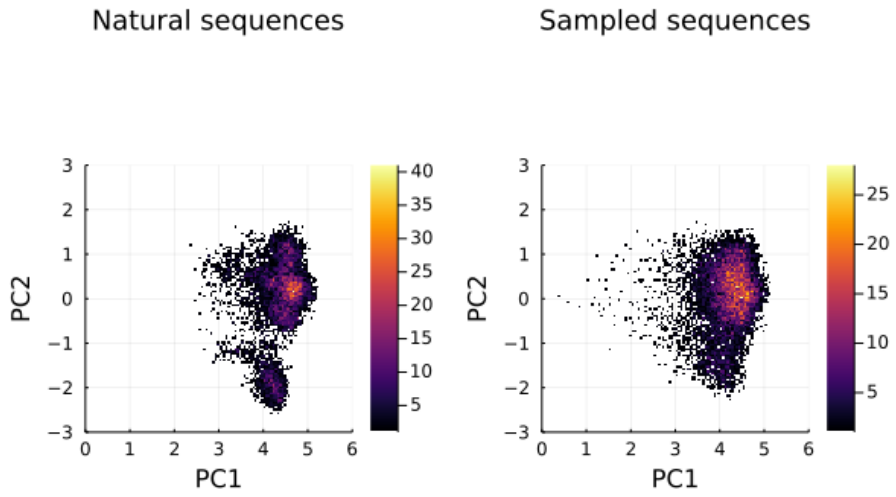

Figure 14: PCA comparison for the PF14 protein family

After showing the main results concerning the artificial sequences generated through the asymptotic method, we can go on by implementing a similar analysis to the contrastive method's ones.
As already mentioned, I implemented the algorithm with a number of sweeps equal to the first five powers of two (including the zeroth power $2^0 = 1$). In some subsequent plots there will be a 0 sweeps sign, meaning that the data is referring to the natural dataset, while the ∞ sign is referring to the asymptotic method results.
Figure 15 shows the value of the two-point connected correlation Pearson coefficient. Trivially for the natural sequences' value we have 1, since it's a comparison between the data with itself, while for the other ones we get approximately 95% of Pearson coefficient, meaning a good similarity with the natural connected correlation, which leads us to believe that the algorithm is working well in predicting the direct contacts.
Another feature of the plot is the slight decrease of the values as the number of sweeps increases: a justification could be that trying to change each residue a number of times more than necessary brings the configuration further from the one it started off, which is indeed a natural sequence.
Similarly to what we just did, we could also compute the three-point connected correlation, excluding this time the natural dataset result from the graph.

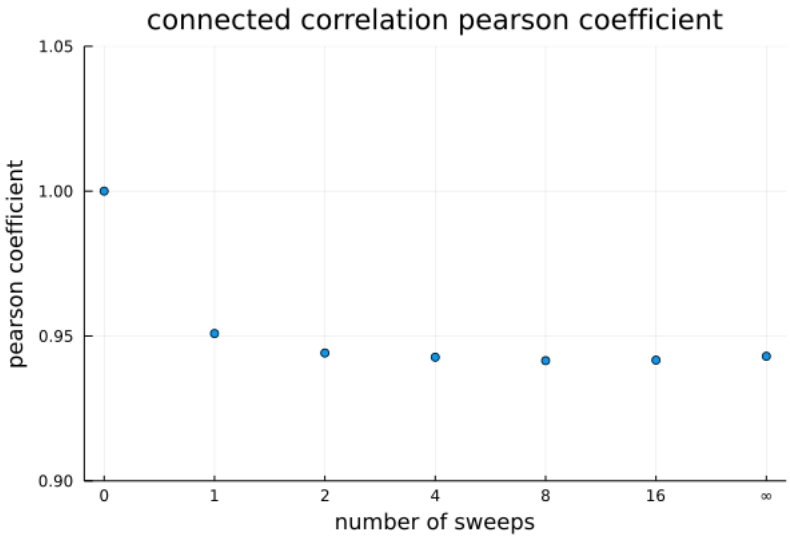

Figure 15: Connected correlation Pearson coefficient of the PF14 protein family

Since the total number of combinations of indices and residues is overwhelming, I decided to derive the connected correlation only on a subset of approximately 10^5 sextets, whose selection method will be clarified in the equivalent paragraph of section 3.2. Looking at the plot shown in figure 16 we see an analogue trend of the coefficient, with respect to the two-point case, as the number of sweeps increases, while its values are, as we should expect, definitely smaller compared to the latter.

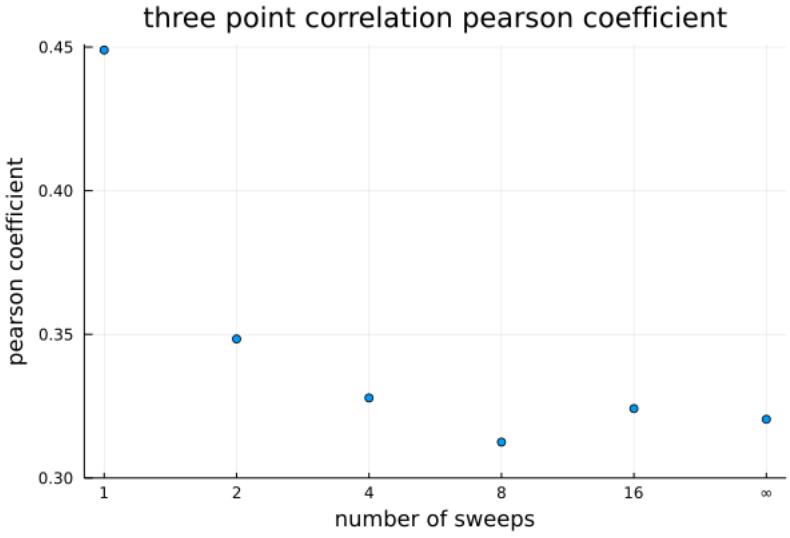

Figure 16: Three-point connected correlation Pearson coefficient of the PF14 protein family

As an alternative to the direct plot comparison of the positive predictive values

curves, we can proceed by showing the number of the first couple of sites for which the precision reaches ~ 0.8, divided by the length of the alignment of the considered family (53 for the PF14 family). A higher number of the variable suggests a greater number of sites with a perfect prediction ($P_{ij} = 1$).

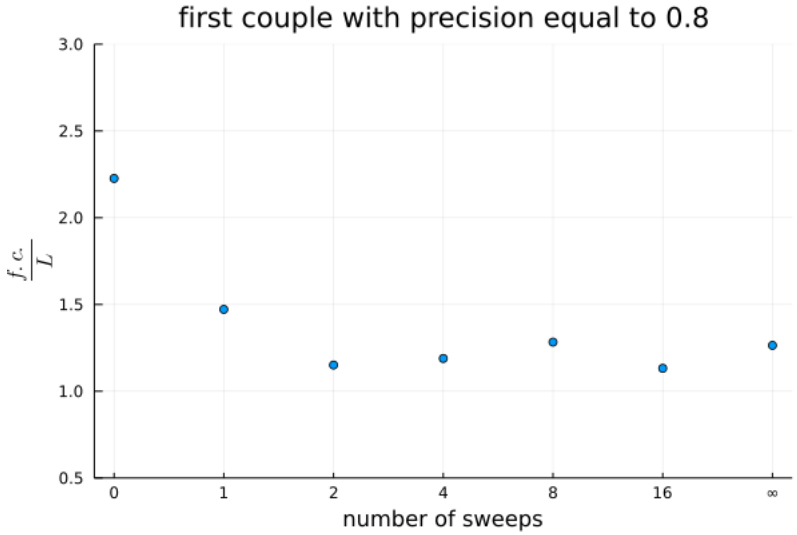

Figure 17: First couple of sites with 80% of precision for the PF14 protein family

While the best value goes not surprisingly to the natural sequences dataset, the other ones, including the asymptotic method result, seems to oscillate around 1.2, with the 1-sweep outcome taking the second place with almost 1.5.
As we will see from other comparisons it is frequently the 1-sweep result that works better according to results, suggesting us that performing too many changes could lead us to a misbehaving sequence. On the other hand we could think that trying only once to flip a residue is not enough, and that the artificial proteins generated through this method would be too much similar to the natural ones we started off.
To settle this doubt, we can give a look at figure 18, where is displayed the mean distance between each generated sequence and the natural one it evolved from, divided by the length of the alignment. We see that despite the considerably lesser number of changes, the 1-sweep result doesn't seem to be far lower than the 16-sweeps' one, and in any case it is higher than 0.5, meaning that more than half of the natural sequence's amino acids changed during the process.
Another similar observable we could look at is shown in figure 19, and it represents the mean of the minimum distances between each generated sequence and the whole natural dataset, divided again by the length of the alignment. Something we should keep in mind is that this minimum distance is not necessarily the beginning natural sequence's, so the plot is usually different from the previous one we have already seen. Sure enough, as we can see from the graph, we get a similar trend but a much lower variation of the values.
To conclude the section with a final result, we can see in figure 20 the comparison among all the PCA histograms referring to the different number of sweeps applied in the contrastive method, plus the previously seen plot (in figure 14) of

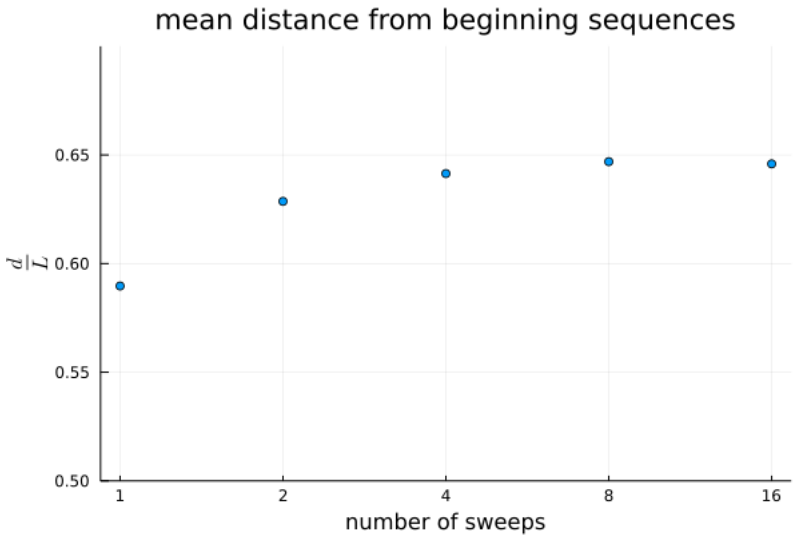

Figure 18: Mean distance from beginning sequences for the PF14 protein family

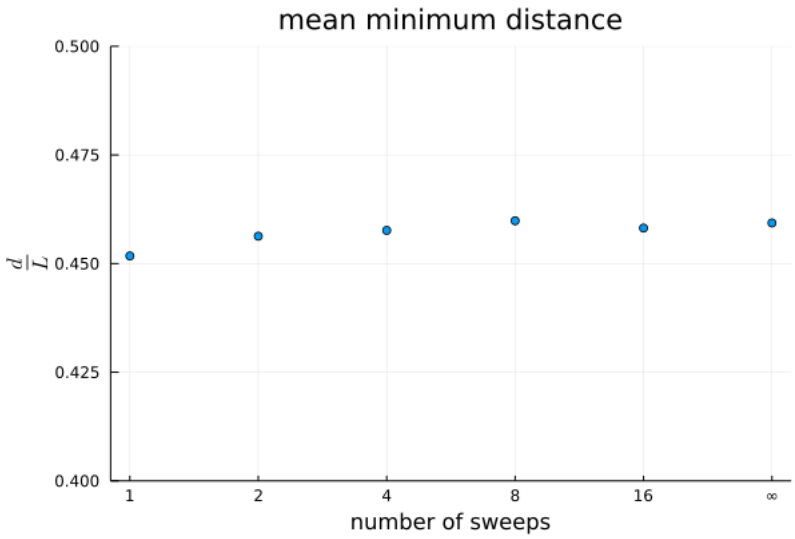

Figure 19: Mean minimum distance from beginning sequences for the PF14 protein family

the natural sequences, included as a reference point. Please note that principal components' labels on the x and y axes (first component for the x axis, second component for the y axis) has been removed from all the individual plots in order to get a good looking figure from the `histogram2d` function of the `Plots` package.
By just a brief look we can see that all the contrastive method's plots are almost identical among themselves and with respect to the asymptotic method's one we have already seen in figure 14, implying that the principal component anal-

ysis is in this case ineffective in detecting a possible difference in performance between the two methods. However the general result is still satisfying, since the shape and color distribution of the plots match sufficiently well with the natural sequence's one.

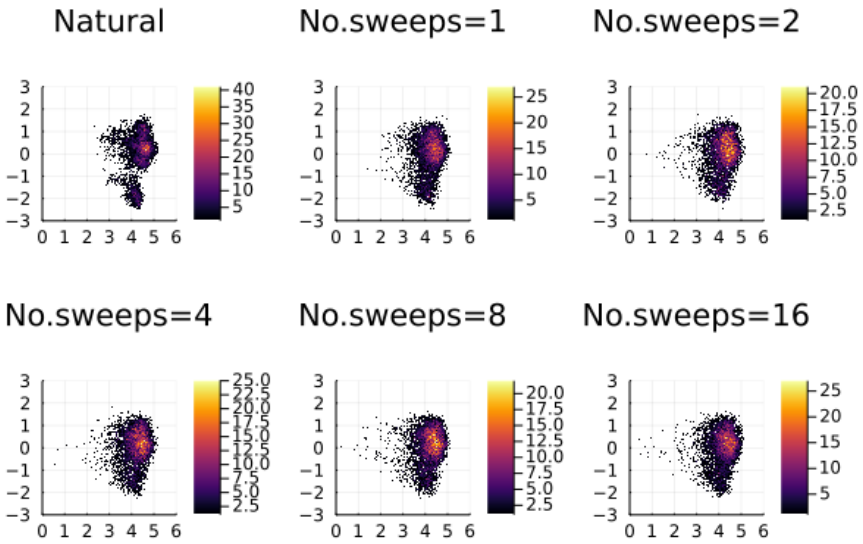

Figure 20: PCA plot for the PF14 protein family

3.2 Other protein families

Beyond analysing the PF14 dataset, I had the chance to work on three other protein families, which are the PF00595, PF00076 and the TEM1 families. For the first two families I will apply a similar abbreviation to the one previously used for the PF14 family (denoting them as PF595 and PF76) for the same reasons.
The PF595 family is composed by 15299 sequences, each one 82 amino acids long. The PF76 family is instead composed by 79366 sequences of 70 residues and at last we have the TEM1 family, with 7515 sequences of 202 residues.
We immediately see that these numbers of sequences and alignment lengths are noticeably different from the ones characterizing the PF14 family, so we already expect the results in these cases to slightly differ from the ones I previously shown.
The first quantity we will address is the connected correlation comparison between the the natural sequences' values (x axis) and the sampled sequences' ones (y axis), generated through the asymptotic method.
Looking at figure 21 we see a substantial linear shape of the two plots concerning protein families PF595 and PF76, similarly to what we have already seen for the PF14 family in figure 12, meaning that the connected correlations of the two datasets are alike. The TEM1 plot however deviates from this behaviour, showing a broader area occupied by the scatter dots. These last observations are mirrored by the Pearson correlation coefficients' estimations, which corresponds to $\rho = 0.92$ for the PF595 protein family, $\rho = 0.96$ for the PF76 and $\rho = 0.80$ for the TEM1.
The inefficiency, related to TEM1 family, in generating sampled sequences with a similar connected correlation to the one of the natural sequences is probably due to the relatively long alignment sequence of its MSA, 202 residues, with respect to the PF14 family's, which was just of 53 residues. This guess is backed up by the data and plots from the PF595 and PF76 families, which have an alignment length of the same order of the PF14's and share the same linear behaviour. Another possible explanation could be the presence of multiple subfamilies in the MSA, which will be pointed out by the PCA plots at the end of the section.
After having derived the connected correlation for the asymptotic method, we now give a look at the Pearson correlation coefficient for the contrastive method in figure 22.
While for the PF595 and PF76 protein families we get an overall good result, with all the coefficients above 0.9 and a slight decrease as the number of sweeps increases in accord to previously seen results, for the TEM1 family instead we see an unexpected peak at No.sweeps=4 with all the values below 0.9, showing the same relative inaccuracy of the asymptotic method.
To conclude our analysis of the connected correlation we can go on representing in figure 23 the three-point connected correlation's Pearson coefficient. Given the fact that the alignment length n of the three families is approximately of the order of 10^2 residues and that we have to consider all possible combinations of sites and amino acids into sextets to compute the connected correlation in its entirety, the time needed to elaborate such a huge number of data would be unsustainable for the hardware in my possession. However, I had the chance to retrieve from the work of Anna Paola Muntoni a more modest number of sextets

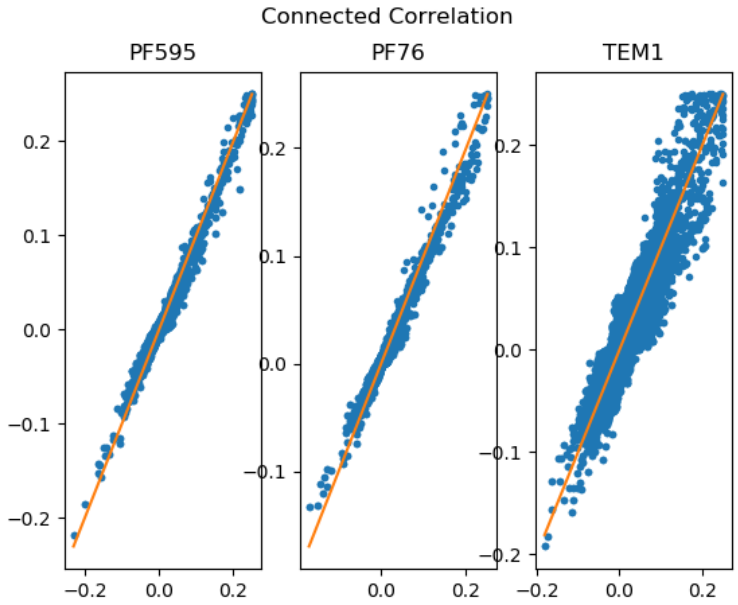

Figure 21: Connected correlation for PF595, PF76 and TEM1 families

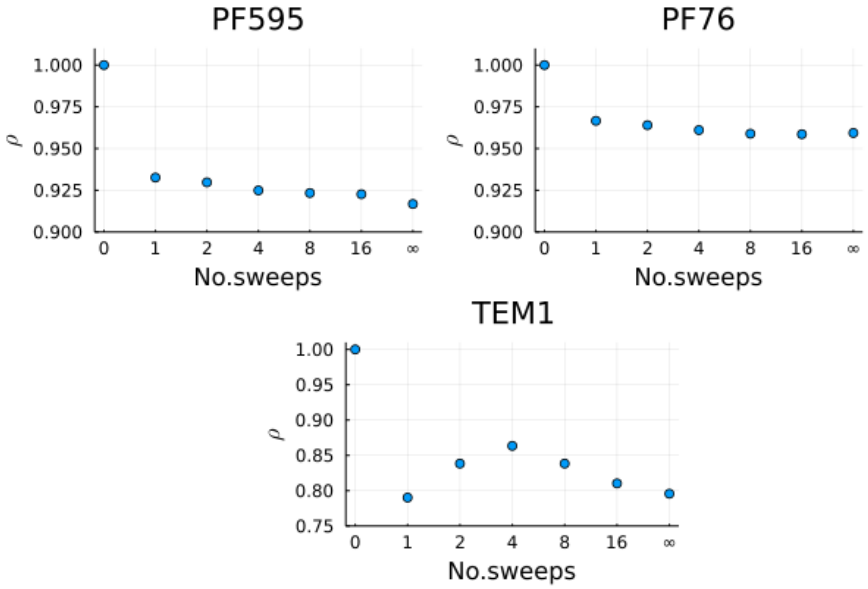

Figure 22: Connected correlation for the contrastive method of PF595, PF76 and TEM1 families

whose connected correlation is guaranteed to be not negligible. Such estimation

was accomplished by calculating the three-point empirical frequency

$$f_{i,j,k}(a,b,c) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(\sigma_i^{(m)}, a)\delta(\sigma_j^{(m)}, b)\delta(\sigma_k^{(m)}, c)$$

for all possible triplets of indices $\{i, j, k\}$ and residues $\{a, b, c\}$. Those that returned a value of empirical frequency above a certain threshold where the ones included in the subset considered for the connected correlation evaluation.
Despite benefiting from this shortened list of indices, I still couldn't derive the connected correlation for the TEM1 family, since, given the either way great number of sextets included in its set, the computer in my possession wasn't able to return an output value in reasonable times. Hence, just for this family I decided to compute the connected correlation over a further subset of indices, which I arbitrarily derived from the former one by selecting one sextet every hundred, ending up with $\sim 3 \cdot 10^5$ different combinations. By looking at the results shown in figure 23, we will have to bear in mind this approximation when considering the TEM1 family plot.
While the outcome for the PF76 family doesn't deviate much from what we have already seen for the PF14's one, the PF595 and TEM1 families instead are significantly worse, with the values floating around 0.10/0.15, although the two-site connected correlation for the PF595 family appeared consistent with the PF76's result.
Another peculiarity concerning the three-point connected correlation of the PF76 family is that the asymptotic method's value this time, on contrary to the general trend of the other plots, outperforms all the contrastive method's ones.

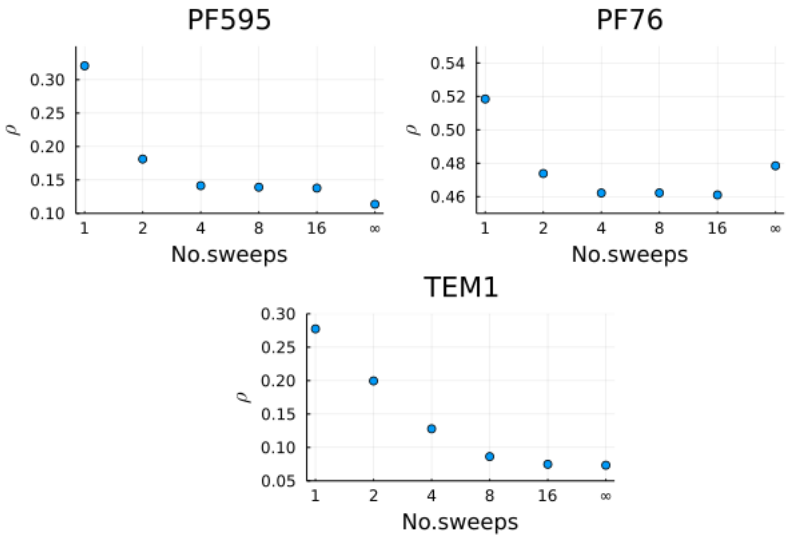

Figure 23: Pearson three-point correlation coefficient for PF595, PF76 and TEM1 families

The next quantity of interest to analyze is the positive predictive value plot of the asymptotic method, which is shown in figure 24, with the already used

limited set of the first 250 couples of sites in the chain with highest scores.
In all of the three plots we can notice that the direct contacts are well predicted by the sampled dataset, with a particularly positive outcome for the PF76 protein family and a slight propensity of the TEM1 sampled curve to diverge from the natural one at higher numbers.

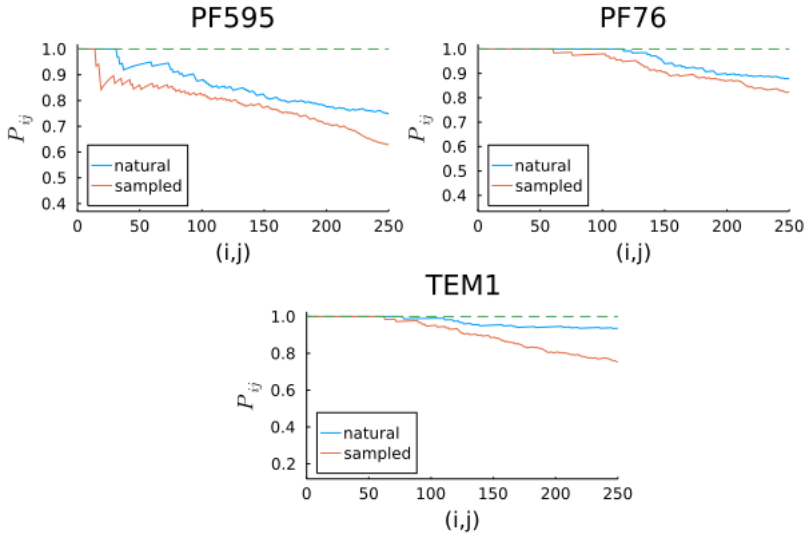

Figure 24: PPV plot for PF595, PF76 and TEM1 families

Proceeding further with the contrastive method, we can check figure 25, showing the first couple's number with precision equal to 0.8, divided by the length of the alignment. The most interesting plot is the PF595's one, since we can notice that the highest value, which corresponds to the best contact prediction, this time is not the natural sequences' one, but instead belongs to the 1-sweep contrastive method, remarking the already told good performance of the generating algorithm for few changes applied in the chain.
Continuing to cover all the quantities of interest already seen for the PF14 family, we can now check the mean distance between each generated sequence and the natural one it evolved from, divided by the length of the alignment.
The results are displayed in figure 26, and for all the three families we get a similar trend to the one observed for the PF14 family, but with higher values.
Note again that the values corresponding to the 1-sweep case are only slightly smaller than the other ones and still above 0.5 for all three families, remarking the fact that the algorithm doesn't require a great number of sweeps to change significantly the residues along the chain from the ones of the natural proteins.

The next quantity to determine is the mean minimum distance already explained in the PF14 section, whose plot is visible in figure 27. Again we see a peak at No.sweeps=4 for the TEM1 family, but on the contrary to the connected correlation case this time we are not dealing with a variable related to the similarity with the natural sequences; somehow we could think it's just the opposite, so I believe the two outcomes to be similar only by chance, although we cannot exclude a relation of some kind.

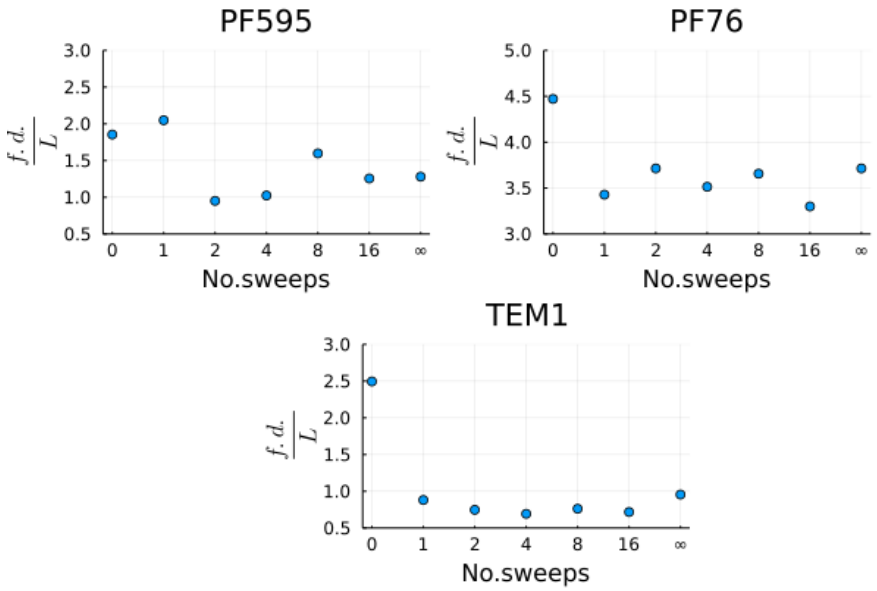

Figure 25: First couple with $P_{ij} = 0.8$ for PF595, PF76 and TEM1 families

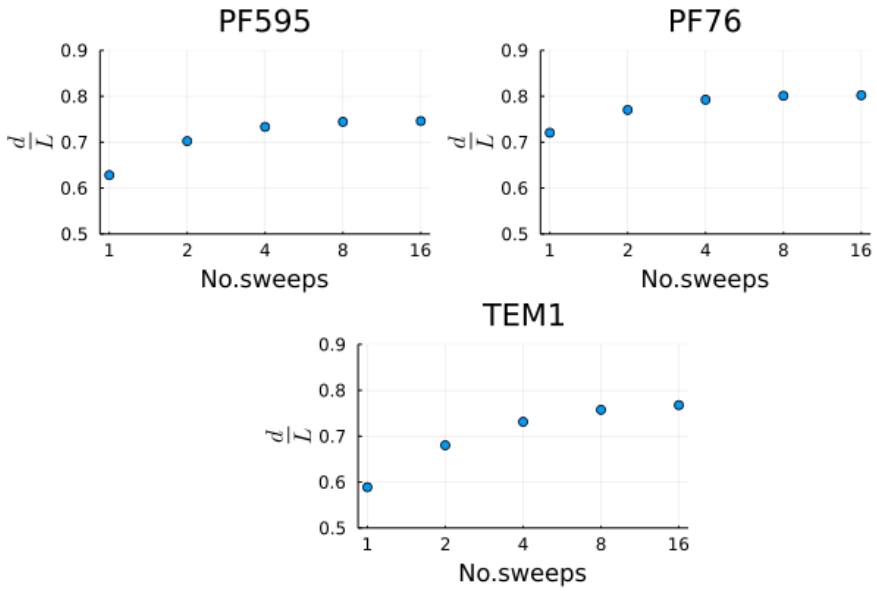

Figure 26: Mean distance from naturals for PF595, PF76 and TEM1 families

To conclude this section we now give a look at the PCA histogram for the three families, where as in the PF14 case I omitted the x-axis label (referring to the first principal component) and the y-axis label (second principal component) for aesthetics reasons.
In order to have a more compact representation of the data, I decided to report in figure 28, aside from the histograms of the natural sequences, only the results

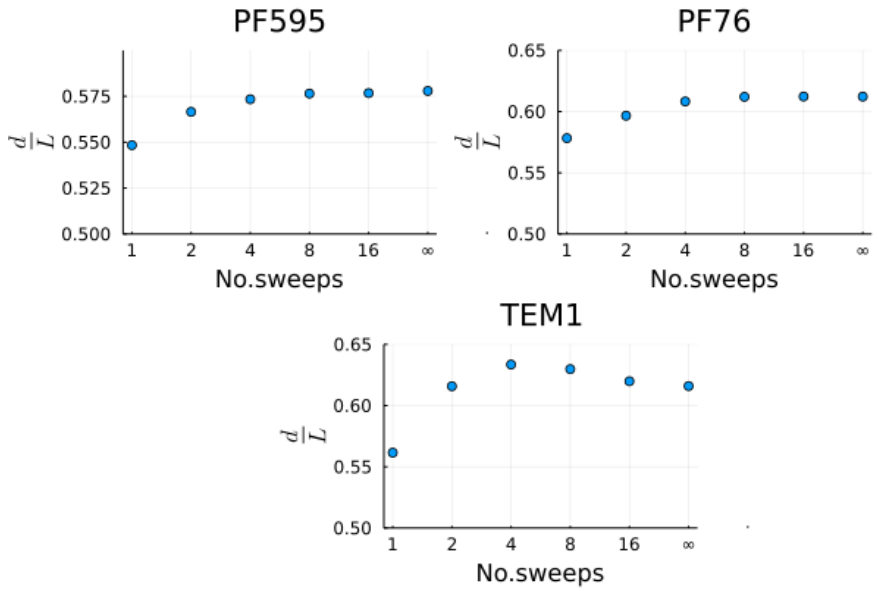

Figure 27: Mean minimum distance from naturals for PF595, PF76 and TEM1 families

regarding the 1, 8 and 16-sweeps methods. The ones that are excluded do not stand out in any particular way from the others, so their omission is acceptable. They are disposed in this way: on the first column from the left we have all the plots regarding the PF595 family, on the second one we have the PF76 family and on the third one the TEM1 family.
We can see from the figure that the shapes and color distributions are alike in all three couples of images, with a surprisingly excellent result for the TEM1 family (especially the 1-sweep method's one), considering the fact that the connected correlation comparison for this family of proteins performed worse than the two others'.
Another peculiarity worth mentioning regarding the TEM1 family is the presence of multiple well-marked islands in the PCA plot of both natural and sampled sequences. As already mentioned in an earlier paragraph in this same section, this type of occurrence can be interpreted as a sign of multiple subfamilies within the same MSA, which could explain the generally worse results obtained for this specific family.

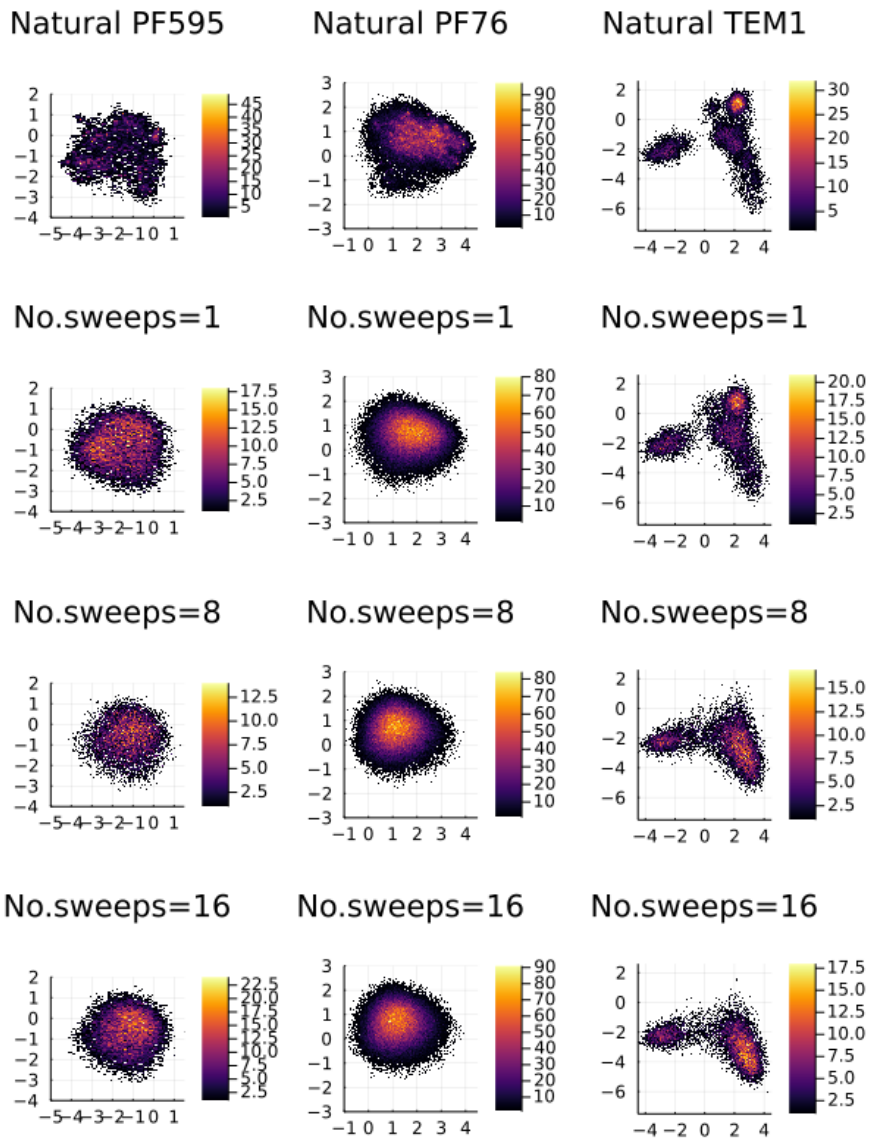

Figure 28: PCA histograms for PF595, PF76 and TEM1 families

4 Conclusion

Looking back at all the data obtained from the two different methods employed and from the four different families of homologous proteins analysed, I personally consider the outcome of this work to be satisfactory. Apart from a possible refinement of the score evaluation method and an improvement in the inference of the parameters of the Potts model as a whole, the Glauber-Pseudolikelihood Dynamics proved to be an efficient technique for the generation of original and feature-reproducing artificial amino acids sequences.
In particular, if I have to identify among the asymptotic method and all the different implementations of the contrastive method a single procedure which proved to be the best one, at least according to the statistical testing tools I considered for the performance evaluation, I would choose the 1-sweep contrastive algorithm.
The outcomes associated with the latter were in many occasions the most positive ones, and despite the modest number of amino acids' changes applied to the sequence, the plots related to the mean distance, such as figure 18 and 26, revealed that the majority of residues in the chain had changed values during the process, implying that the sampled sequences were indeed original.
Looking back at the setting of this thesis work, the consistency between the attributes of the natural and sampled sequences and the originality of the latters were actually the two main goals I wished to achieve for my algorithm, and it seems that the contrastive method met perfectly these prerequisites.
Nevertheless, the asymptotic method did also provide reasonably good results, so I think it should also be considered for an implementation of a problem of this kind. A interesting fact to point out is that the data associated with this last method matched pretty well with the one of the 16-sweeps contrastive method, validating the hypothesis that the asymptotic method's results could be also seen as a long time, or equivalently for a great number of sweeps applied, outcome of the former method.
This assumption was taken as a consequence of the Markov chain set up, in which we prefigured our sequences, both natural and sampled, to be part of an equilibrium subset of the dynamic. The transition probabilities, which corresponded to the pseudolikelihood distribution with the inferred parameters, were indeed the ones set to "lead" the initial random sequences of the asymptotic method to this subset, in which the initial sequences of the contrastive one, i.e. the natural proteins, already belonged to.
Regarding the contrastive method, a key point in the understanding of the latter was to verify if the pseudolikelihood (or energy, if we reason in a more statistical physics' way) minimization criterion used for the transition probabilities of the Markov chain restrains our sampling dynamics inside the desired equilibrium subset, or if it leaves it as soon as the number of residues' changes applied increases. Bearing in mind that the "borders" of this hypothetical subset are not well defined, since we don't know a priori all the sequences belonging to it (otherwise, all the work done would be meaningless), the quantities of interest analyzed revealed, as already pointed out, a worsening of the results at higher sweeps, which however proved to be moderate. This last observation suggests us indeed that, although not perfectly fitting, the theoretical basis underlying the generative model are adequate for the purpose of this project.
One last thing which needs to be underlined is that, due the time in my posses-

sion to carry on this thesis work, only a small number of protein families were taken into consideration. Such limitation opens up to future research about this subject, where, in addition to considering a wider number of MSA, also an implementation of further testing tools could corroborate or disprove the assumptions and assessments present in this work.
Adding to these remarks, the Glauber-Pseudolikelihood Dynamics would surely benefit from a comparison with other generative models, in order to ensure if the pseudolikelihood maximization principle behind this work is effectively a good premise for the solution of the artificial proteins generation's problem, which remains to this day a key subject in biology applications.

5 Acknowledgements

As this thesis work comes to an end, I would like to make use of this last chapter to express my sincere gratitude towards all the people who assisted me during this last project and throughout my whole university course.
First of all, I would like to thank my family, who supported me in this journey from the beginning and without whom I wouldn't have the possibility of accomplishing this result. I would like to extend my gratitude to Cecilia and all my friends, who filled my days with joy and sustained me in times of need.
Then, I would like to acknowledge the great help of my supervisors, professor Andrea Pagnani and Anna Paola Muntoni, who guided me in this research and provided me all the assistance I needed.
Lastly, a special thanks to my grandma, who wished so hard to see me graduate and left me with tender memories.

A Kullback-Leibler divergence

Given two probability distributions $p(x)$ and $q(x)$ defined over a discrete space X, the Kullback-Leibler divergence [17] between the latters is defined as

$$KL(p||q) = \sum_{x \in X} p(x) log \left(\frac{p(x)}{q(x)} \right) \tag{25}$$

Obviously for such an equation to make sense, we will have to assume that $q(x) \neq 0 \ \forall x \in X$.
Thanks to Jensen's inequality [15], it is possible to show that $KL(p||q) \geq 0$ for all probability functions $p(x)$ and $q(x)$, with the 0 value achieved only in the case $p(x) = q(x) \ \forall x \in X$. Keeping in mind this last consideration, we could interpret the Kullback-Leibler divergence as a measure of how much different the two probability distributions are.
If we choose in particular the $q(x)$ function to be equal to the uniform distribution $U(x) = \frac{1}{|X|}$ over space X, we get

$$\begin{aligned} KL(p||U) &= \sum_{x \in X} p(x) log \left(\frac{p(x)}{|X|^{-1}} \right) \\ &= \sum_{x \in X} p(x) log(p(x)) - log(|X|^{-1}) \sum_{x \in X} p(x) \\ &= -S(p) - log(|X|^{-1}) \end{aligned} \tag{26}$$

We easily see from this last result that minimizing $KL(p||U)$ over the whole space of distributions $p(x)$ (so, as we have learned, finding among this set of functions the one more similar to the uniform distribution) is equivalent to maximizing the entropy $S(p)$, which indeed can be interpreted as the degree of "flatness" of $p(x)$.
If we now consider two probability distributions $p_X(x)$, $p_Y(y)$ and their joint distribution $p_{(X,Y)}(x, y)$, we are able to define the Kullback-Leibler divergence between the latter and the products of the two marginals, which takes the name of mutual information (MI)

$$\begin{aligned} MI(X, Y) &= KL(p_{(X,Y)}(x, y)||p_X(x)p_Y(y)) \\ &= \sum_{x \in X} \sum_{y \in Y} p_{(X,Y)}(x, y) log \left(\frac{p_{(X,Y)}(x, y)}{p_X(x)p_Y(y)} \right) \end{aligned} \tag{27}$$

From the properties of the Kullback-Leibler divergence, the Mutual Information is still a non-negative quantity, and this time the 0 value is obtained only in the case $p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$, so in other words, if the two random variables X and Y are statistically independent. Given this fact, we can use the mutual information in order to establish how much two events are mutually influenced. Applying this concept to the couplings' identification inside a protein chain, we could try to define a site-to-site mutual information of this type

$$MI_{ij} = \sum_{a,b} f_{ij}(a,b) log\left(\frac{f_{ij}(a,b)}{f_i(a)f_j(b)}\right) \tag{28}$$

where $f_{ij}(a,b)$ and $f_i(a)$ are the usual empirical frequency functions referring to a multiple sequence alignment of homologous sequences

$$\begin{cases} f_i(a) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(a, \sigma_i^{(m)}) \\ \\ f_{ij}(a,b) = \frac{1}{M_{eff}} \sum_{m=1}^{M} \omega_m \delta(a, \sigma_i^{(m)}) \delta(b, \sigma_j^{(m)}) \end{cases} \tag{29}$$

The following step would be to determine this quantity for all possible couples of sites in the sequence and then set a threshold for MI_{ij} above which the pair (i,j) represents a contact in the structure. However the problem of differentiating direct contacts from indirect ones makes this approach inadequate. A possible solution could be modifying the mutual information into

$$MI_{ij} = \sum_{a,b} P_{ij}^{dir}(a,b) log\left(\frac{P_{ij}^{dir}(a,b)}{f_i(a)f_j(b)}\right) \tag{30}$$

where the direct contact distribution $P_{ij}^{dir}(a,b) \sim e^{J_{ij}(a,b)+h_i(a)+h_j(b)}$ satisfies the maximum entropy principle already discussed in section 2.1, with the parameters $\{J_{ij}(a,b), h_i(a)\}$ being the Lagrange multipliers of this problem.

B Data block division

The data block division is an useful method for dealing with a large number of information. It essentially enables you to recompact your data and at the same time extract its mean value and variance behaviour.
It operates implementing division of your strings or vectors of data in subsequently shorter halves and extract from all these "fragments" their mean value and variance. To be more clear: having an initial vector of length n, the first division generates the first subvector containing entries of the original one in the interval $[\frac{n}{2}+1, n]$, and then we proceed by calculating the mean value and variance μ_1, σ_1^2 of the latter. Then we take the remaining subvector with entries $[1, \frac{n}{2}]$ and apply the same procedure, furtherly dividing it in two halves and then calculating μ_2, σ_2^2 of the subvector with entries $[\frac{n}{4}+1, \frac{n}{2}]$ of the original vector. We implement this operation until exhausting all the elements in the vector, and then plot the result through the `errorbar` function of the `PyPlot` library of Julia.
Here an example of the result of the procedure: after generating 1000 numbers between 0 and 1 with a normal gaussian distribution, we report on the left of figure [29] the whole data, and on the right its block division.

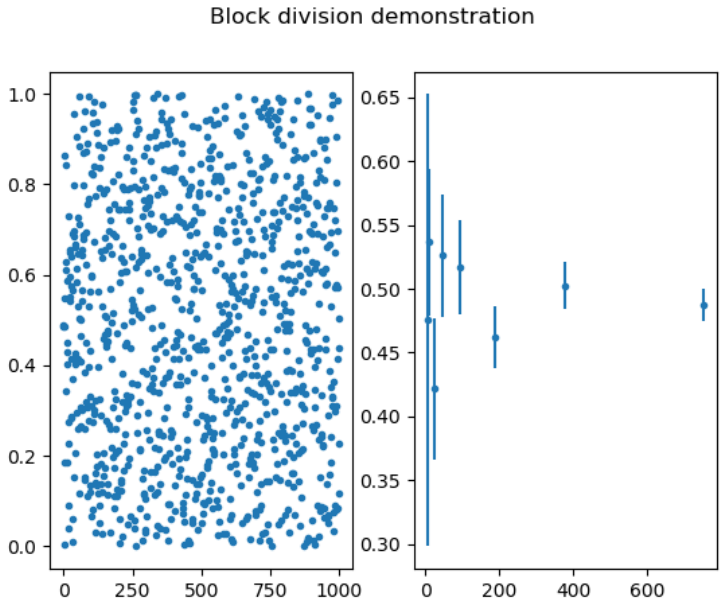

Figure 29: Data (left) and their block recompacting (right)

Bibliography

[1] J. Lewis B. Alberts A. Johnson. *Molecular Biology of the Cell*. Garland Science, 2015.

[2] A.J.W. te Velthuis C.P. Bakowski W. Bruins. "The Nature of Protein Domain Evolution: Shaping the Interaction Network". In: *Current Genomics* 11(5) (2010), pp. 368–376. DOI: 10.2174%2F138920210791616725.

[3] M. Á. Carreira-Perpiñán and G. E. Hinton. "On Contrastive Divergence Learning". In: *AISTATS* (2005).

[4] G. M. Cooper and R. E. Hausman. *The Cell: A Molecular Approach*. Sinauer Associates Inc, 2013.

[5] T. Taimre D. P. Kroese T. Brereton and Z. I. Botev. "Why the Monte Carlo method is so important today". In: *WIREs Computational Statistics* 6(6) (2014), pp. 386–392. DOI: 10.1002/wics.1314.

[6] C. Lövkvist E. Aurell M. Ekeberg, Y. Lan, and M. Weigt. "Improved contact predictions in proteins: Using pseudolikelihoods to infer Potts models". In: *Physical Review* E87.012707 (2013). DOI: 10.1103/PhysRevE.87.012707.

[7] B.Lunt F. Morcos A. Pagnani et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proceedings of the National Academy of Sciences* 108(49):E1293–E1301 (2011). DOI: 10.1073%2Fpnas.1111471108.

[8] K. Bailey F. Sanger. "The chemistry of amino acids and proteins". In: *Annual Review of Biochemistry* 20 (1951), pp. 103–130. DOI: 10.1146/annurev.bi.20.070151.000535.

[9] P. A. Gagniuc. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons, 2017.

[10] R. J. Glauber. "Time-Dependent Statistics of the Ising Model". In: *Journal of Mathematical Physics* 4,294 (1963). DOI: 10.1063/1.1703954.

[11] H. v. Helmholtz. *Physical memoirs, selected and translated from foreign sources*. Taylor & Francis, 1882.

[12] L. Williams J. Mistry S. Chuguransky et al. "Pfam: The protein families database in 2021". In: *Nucleic Acids Research* 49 (), pp. D412–D419. DOI: 10.1093/nar/gkaa913.

[13] J. Gajl-Peczalska J.H. Kersey. "T and B lymphocytes in humans. A review." In: *The American Journal of Pathology* 81(2) (1975), pp. 445–458.

[14] R. Jacobsen. "Artificial Proteins Never Seen in the Natural World Are Becoming New COVID Vaccines and Medicines". In: *Scientific American* (2021).

[15] J. L. W. V. Jensen. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". In: *Acta Mathematica* 30(1) (1906), pp. 175–193. DOI: 10.1007/BF02418571.

[16] I. T. Jolliffe. *Principal Component Analysis*. Springer New York, 1986. DOI: 10.1007/978-1-4757-1904-8_7.

[17] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *Annals of Mathematical Statistics* 22(1) (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

[18] Zhang lab. *What is FASTA format?* URL: https://zhanggroup.org/FASTA/.

[19] T. Hartonen M. Ekeberg E. Aurell. "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences". In: *Journal of Computational Physics* 276 (2014), pp. 341–356. DOI: 10.1016/j.jcp.2014.07.024.

[20] A. Veglio M. Semplice A. Gamba, G. Naldi, and G. Serini. "A Bistable Model of Cell Polarity". In: *PLOS ONE* 7(2):e30977 (2012). DOI: 10.1371/journal.pone.0030977.

[21] T. M. Mitchell. *Machine Learning*. McGraw-Hill Education, 2019.

[22] A. Krogh R. Durbin S.R. Eddy. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 2002.

[23] M. Figliuzzi S. Cocco C. Feinauer, R. Monasson, and M. Weigt. "Inverse Statistical Physics of Protein Sequences: a Key Issues Review". In: *Reports on Progress in Physics* 81.032601 (2018).

[24] M. S. Majik U. B. Gawas V. K. Mandrekar. "Advances in Biological Science Research: A Practical Approach". In: Academic Press, 2019. Chap. 5.

[25] F. Y. Wu. "The Potts model". In: *Reviews of Modern Physics* 54(1) (1982), pp. 235–268. DOI: 10.1103%2FRevModPhys.54.235.