Politecnico di Torino

Master’s Degree in Physics of Complex Systems

Academic Year 2023/2024

Master’s Degree Thesis

Analysis of protein families using a reduced space representation and the Expectation Propagation method

Supervisors:

Andrea PAGNANI

Anna Paola MUNTONI

Candidate:

Alessandro Macis

Summary

The main focus of this thesis is the description, and then the application, of a method involving dimensionality reduction in order to analyse different families of homologous proteins.

The dimensionality of the space in which the homologous proteins under study live varies, depending on the specific protein family, in particular, it depends on the number of possible different amino acid residues (q) and the length of the protein sequences (L). In the course of this study we analyse families with the same value of q (q=21) and varying L.

To simplify the analysis, the protein sequences are one hot encoded, that is, each of the amino acid residues that constitute the protein is expressed as a sequence of q numbers, all of which are zeros except for one 1 placed on a different position along the q-length sequence depending on the type of starting residue. The resulting vector is then only composed of zeros and ones and has dimension qL.

The dataset we are working with is structured as a Multiple Sequence Alignment, which can be seen as a matrix containing homologous proteins, belonging to a given protein family, in its rows. The matrix representing the MSA has dimensions $M \times L$, with M being the number of sequences in the family and L the aforementioned length of those sequences. After the process of one hot encoding, the MSA matrix has instead dimensions $M \times qL$.

To perform the reduction of dimensionality mentioned in the beginning of the paragraph we use Principal Component Analysis (PCA), this method allows us to map the qL dimensional data contained in the MSA onto an N_c dimensional space, identified by the first N_c principal components, while minimising the loss of information due to the projection. The principal components are computed from the dataset and the number of dimensions of the reduced space (N_c) can be chosen to best fit our needs (eg. a graphical representation of the results).

The value of N_c can be modified by retaining a higher or lower number of eigenvectors of the covariance matrix, which are shown to be the principal components.

Together with PCA, the Expectation Propagation method is also used, with the aim of estimating the probability distribution of a sequence that fits the projection constraints imposed with the use of PCA. the application of the EP method in this context is, in fact, to find the protein sequence corresponding to a given projection in the N_c-dimensional space, effectively solving the inverse problem with respect to the projection performed via PCA.

The utility of the combined use of both the PCA and Expectation Propagation methods comes from the enhanced interpretability (eg. graphical interpretability) of a lower dimensional space, which is also generally easier to work in, while hopefully retaining the important biological information contained in the original dataset. Another notable advantage of the use of PCA is the reduction of the noise of the problem. The advantage of the use of the Expectation Propagation method is that, owing to its ability to estimate the probability of a sequence associated to a specific projection, it provides a way through which we

can better interpret, in the sense of biological interpretation, results obtained in the reduced dimensional space.

To sum it up, by using PCA and EP to create a two-way mapping between the full dimensional and the N_c-dimensional space, we are able to take advantage of the aforementioned positive features coming from the reduction of dimensionality while still retaining a certain degree of biological interpretability in the results.

The deciding factor in determining the effectiveness of the method is then the quality of the mapping, this, together with the results of various analysis in the reduced dimensional space, is reported in the thesis.

We now summarise some of the results considered in the study.

The analysis of the Potts energy landscape (used to model the fitness of the sequences) in the N_c-dimensional space leads to a structure formed by wells positioned in correspondence of the projections of the natural sequences and curves with higher energy values in the space between the sequences.

The computation of the variation of the Hamming distance, following a displacement on the reduced dimensional space, confirms the intuitive idea of an approximately linear increasing relation between the two quantities. This kind of relation applies, in particular, to relatively small displacements in the neighborhood of a starting sequence, as, when the curve approaches the maximum possible value (the length of the sequence) its behaviour changes.

Finally, results of statistical analysis performed on sets of sampled sequences show how the method is actually able, despite the approximations, to retain the biological information contained in the original dataset.

Contents

Chapter 1

Introduction to Principal Component Analysis (PCA)

In this chapter we give a general introduction to the Principal Component Analysis method (PCA), which will be at the core of our proposed analysis presented in the next chapters. This method is first introduced as a way to treat datasets of large dimensions, via the application of PCA. It is in fact possible to increase the interpretability of the dataset, by reducing its dimensionality, while minimising the loss of information due to the transformation. Notice how, when talking about the dimensions of the dataset, we mean the number of components of the data itself and not, for example, the number of repeated experiments producing that piece of data. [4]

The solution for this problem is to look for a new set of uncorrelated variables obtained as linear functions of the original variables, the obtained directions will be called principal components (giving the name to the method). To account for the minimisation of the loss of information we equivalently impose that the new variables preserve as much as possible the variability of the dataset.[6] The first principal component for example can in fact be defined as the direction that maximises the variance of the projected data, the second principal component is associated with the second highest variance, and so on. The resulting variables will constitute a new basis using which we can represent the dataset in a new, lower dimensional, space.[7]

Variables that satisfy such conditions will be obtained as solutions to an eigenvalue problem, as shown in the next paragraph. [1]

1.1 Computation of Principal Components

We now briefly show how the condition of uncorrelated variables maximising the variance can be satisfied by solving an eigenvalue problem.

We start by more formally presenting the context of the approximation. The starting dataset $\mathbf{X}$ is in the form of an $M \times L$ matrix, where the M rows can be interpreted as different realisations of an experiment, and the L elements of the columns can be seen for example as different quantities observed in the course of each experiment, the reduction of dimensionality operated by PCA applies to the latter (L). The specific interpretation of M and L in our case study will be presented in the following chapters.

The next step is the introduction of a transformation mapping each row vector $\mathbf{x}_i$ to a new vector $\mathbf{t}^i$ of size p. Generally, when applying the PCA method, p is chosen to be smaller or much smaller than L, a more in-depth discussion about this can be found in the next section, about the general properties of PCA.

Said transformation can be expressed in terms of a set of vectors $\mathbf{w}_j$ of length L, with $j = 1, \dots, p$. The relation between the vectors $\mathbf{w}_j$ and the principal components is the following:

$$t^i_j = \mathbf{x}_i \cdot \mathbf{w}_j \qquad\qquad i = 1, \dots, M \quad , \quad j = 1, \dots, p$$

Our aim is to find the set of $\mathbf{w}_j$ such that the variance of the dataset along the principal directions is maximised. [2] We also impose that $\mathbf{w}$ must be an unit vector.

We can write the first vector $\mathbf{w}_1$, which will then identify the first principal component, satisfying the aforementioned conditions as:

$$\mathbf{w}_1 = \arg \max_{||\mathbf{w}||=1} \sum_{i=1}^{M} (t^i_1)^2 = \arg \max_{||\mathbf{w}||=1} \sum_{i=1}^{M} (\mathbf{x}_i \cdot \mathbf{w}_1)^2$$

By reintroducing the matrix $\mathbf{X}$ representing the dataset

$$\begin{aligned} \mathbf{w}_1 = \arg \max_{||\mathbf{w}||=1} ||\mathbf{X}\mathbf{w}_1||^2 = \arg \max_{||\mathbf{w}||=1} (\mathbf{w}_1^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1) = \\ = \arg\max(\frac{\mathbf{w}_1^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1}{\mathbf{w}_1^T \mathbf{w}_1}) \end{aligned} \tag{1.1}$$

The term in parenthesis in the last step of the equation is called Rayleigh quotient, and is maximised if $\mathbf{w}_1$ is the first eigenvector of $\mathbf{X}^T\mathbf{X}$. [8]

We note how the product $\mathbf{X}^T\mathbf{X}$ can be interpreted as the empirical covariance matrix, for the application of the PCA method to our case of study we will in fact use the covariance matrix computed from the dataset.

It can similarly be shown that the remaining principal components are obtained by choosing $\mathbf{w}_j$ equal to the j-th eigenvector of $\mathbf{X}^T\mathbf{X}$ (of the empirical covariance matrix).

The obtained results can be summarised as:

$$\mathbf{T} = \mathbf{X}\mathbf{O}$$

where $\mathbf{O}$ is the matrix whose columns are the eigenvectors of $\mathbf{X}^T\mathbf{X}$. $\mathbf{O}$ is then a $L \times p$ matrix.

1.2 General properties

Notably, the results of the application of the PCA method only depend on the dataset itself, in the sense that there is no prior distributional assumptions made about the variables we are looking for. Coherently with this approach, the Principal Component Analysis method can be said to be a descriptive method instead of an inferential one, in this sense PCA can be applied to very different kinds of problems and different types of data, being able to give clearer insight on the information encoded in the dataset.

An important note has to be made regarding the choice of the number of principal components which will be taken into consideration. In terms of graphical interpretability of the results the obvious choice is to take the first two principal components (again, as "first two" we mean the two directions associated with the highest variance), by doing so it is possible to visualise the projections of the elements of the dataset on a plane. [3] This kind of visualisation can be particularly meaningful in this setting, owing to the choice of principal components as directions that maximise variance, in fact, eventual clusters of data projections will be maximally spread out, facilitating their recognition. Instead, if we choose two random directions and visualise the projections of the dataset on that plane, the cluster may substantially overlay, making them indistinguishable.

As said at the start of paragraph 1.1, the number of columns of the matrix **O** (p), which corresponds to the number of retained eigenvectors of the covariance matrix, is usually chosen to be smaller than L. In this case, the projection onto the p-dimensional space is an approximation of the original dataset. In the choice of the number of retained components we can also take into consideration a measure of the quality of this reduced dimensional approximation. A possible measure is the variability associated with the set of retained principal components, that is the fraction of variance which is explained by the chosen set. Explicitly writing the form of the fraction of explained variance for component i:

$$EV_i = \frac{\lambda_i}{\sum_{i=1}^{M} \lambda_i}$$

Given this kind of measure, one can choose to pick a set of principal components such that the quality of the approximation is over a certain threshold or, if the number of components is limited by the need for a graphical visualisation the explained variance can provide an estimation of the quality of the results.

Graphs for the explained variance regarding our subject of study will be presented in the next chapters.

Chapter 2

Biological introduction

In this section, we will present an introduction to the main biological aspects that will be treated in this study, in particular sec. 2.1 briefly describes the structure of proteins and their relevance in biological processes, section 2.2 regards an introduction to Multiple Sequence Alignments (MSA) and finally, in 2.3 we will discuss the derivation of the Potts model form which will be used throughout this study to assess the fitness of a given sequence.

2.1 Proteins

Proteins are one of the most fundamental building blocks for the biological processes of any living being. Their function widely varies, including acting as catalysts in a large number of processes (enzymatic function), giving structure to cells (cytoskeleton), and transporting molecules, among others. [15] [16]

In agreement with this huge range of functions, we find many different kinds of proteins, differing from each other in structure and composition. The information necessary to "build" each of those proteins is coded in the genes. Despite all of these differences, all proteins can be said to share a general structure. Every protein is in fact a sequence of amino acids [9] [10], organic compounds containing amino($-NH_2$) and carboxylic acid ($-COOH$) groups in addition to which we also find a side chain that differentiates the hundreds of different known amino acids. Among those, 22 are "proteinogenic", that is, they constitute the proteins found in every living being, we also specify that only 21 of those amino acids are found in eukaryotes (20 are normally contained in the genetic code, the other one is obtained via a special translation mechanism).

In the structure of the proteins, amino acids are bound together by a peptide bond, that is a covalent bond linking the carbon of the carboxylic acid of one amino acid and the nitrogen of the amino group of the other. Multiple amino acids linked together create a polypeptide, which effectively constitutes the backbone of the protein, with the side chain specific to each amino acid positioned on the side of this backbone, as shown in figure 2.1

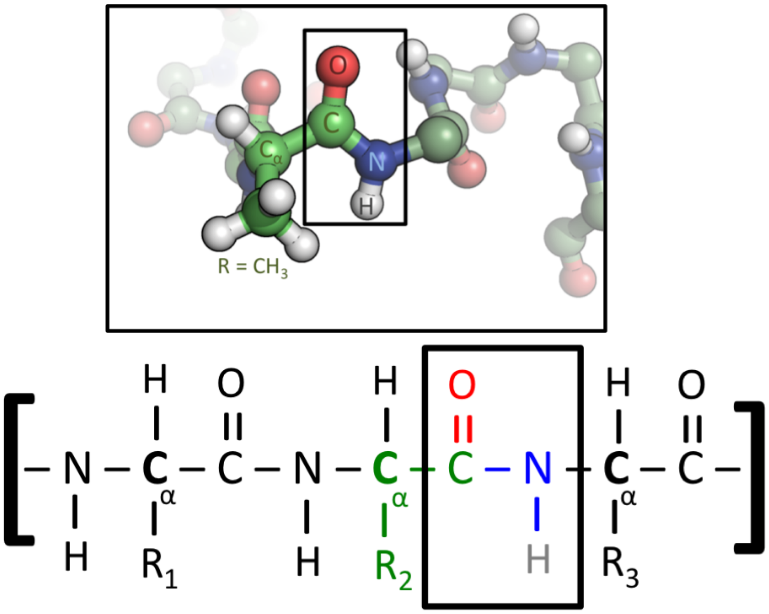


Figure 2.1: Illustration of the primary structure of the protein, image taken from https://en.wikipedia.org/wiki/Protein

Different amino acids can also be classified according to their properties, taking into consideration factors such as hydrophobicity/hydrophilicity, charge, and size. Those properties influence and determine important characteristics such as protein structure and protein-protein interaction. [11] [12]

In particular, we can expand on the concept of protein structure. Probably the most important aspect that needs to be noted about this is the link between the structure and the function of a protein, the possibility to partially infer the utility and the role of a protein in a specific process by observing its structure is enough to generate interest around its analysis.

The structural organization in proteins can be in general divided into different layers:

- *Primary structure*: consists in the actual sequence of amino acids that forms the backbone of the protein
- *Secondary structure*: local structures stabilized by hydrogen bonds, most commonly $\alpha - helices$ and $\beta - sheets$
- *Tertiary structure*: The 3D conformation of the protein, which depends on how the protein folds on itself due to long range interaction between amino acids, this structure thus strongly depends on the chemical and physical properties of the amino acids constituting the chain. The folded state of

the protein is called native state, some proteins fold into their native state autonomously while others need so called molecular chaperones.

The tertiary structure of the protein in particular is strongly related to its function, thus the problem of predicting the 3-dimensional folding of a given protein is widely studied

- *Quaternary structure*: structure formed by different protein molecules which are linked together via inter protein interactions forming a protein complex

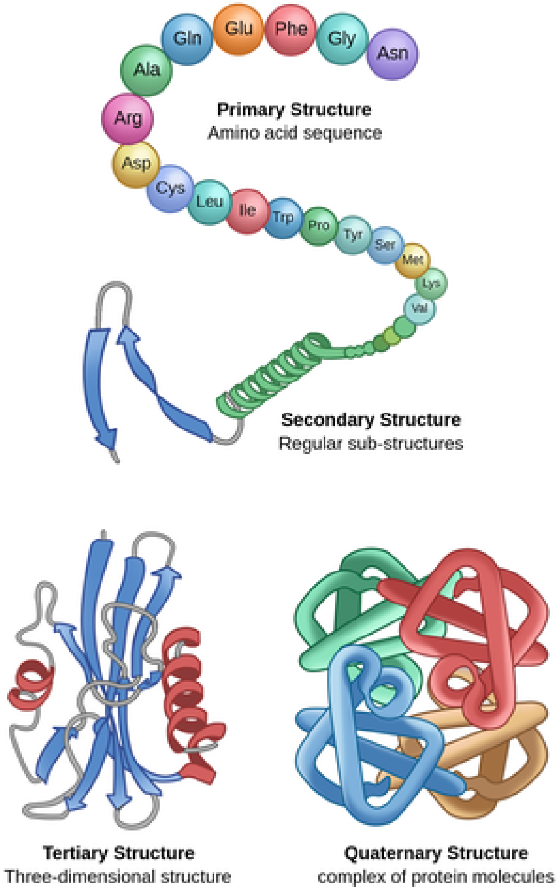


Figure 2.2: Representation of the different levels of protein structures introduced in 2.1. Image taken from https://theory.labster.com/protein-structure/

2.2 Multiple Sequence Alignments

With time, proteins undergo changes in their structure due to mutations of various natures, this can result in the existence of sequences that are different from each other but share a common ancestor, those sequences are said to be homologous [21]. In the context of the alignment, differences among homologous proteins are modeled through three possible moves: insertion, deletion and substitution of single amino acids. The effect resulting from the applications of many of these moves to a starting sequence during the course of time is the difference we see between sequences of homologous proteins. Notice how performing an insertion or a deletion on a protein sequence varies its length, to be able to always compare sequences of the same length we thus need to introduce a gap "-".

Still, homologous proteins generally have similar structures, consequently similar functions, and can be thought to constitute a protein family. We can interpret these homologous proteins as different realizations of a single protein, this intuition of course suggests the possibility to employ statistical methods in the analysis of the protein families.

The data structure that follows from this idea and allows for the application of different methods of analysis to a protein family is the Multiple Sequence Alignment (MSA) [17], which can be modeled as a matrix in which each row is a protein sequence. More formally we can say that the matrix has dimensions $M \times L$, where M is the number of sequences in the alignment and L is their length, and each row can be written as a vector $\mathbf{S}_i = (s_1^i, s_2^i, \dots, s_L^i)$, $i = 1, \dots, M$. Each s_j^i lives over a discrete alphabet of symbols, each representing an amino acid. Generally the number of symbols is 21, 20 canonically coded(?) amino acids, to which we add one symbol encoding the gap, those symbols can be mapped to integer numbers $1, \dots, q$, with $q = 21$ as just stated. In this case the MSA will be represented as a $M \times L$ matrix with numeric entries.

There exist different methods through which the actual alignment of the protein sequences (rows of the MSA) is constructed [18] [19], our aim is in general to find an alignment that maximises some score. Those methods will not be discussed in this introduction as it goes beyond the scope of this work.

We now introduce two quantities, which will be treated in this work in the following chapters, deriving from basic statistical analysis performed on the Multiple Sequence Alignment, and discuss their relevancy regarding biological information.

$$\begin{cases} p_i(a) = \frac{1}{M} \sum_{m=1}^{M} \delta(a, s_i^m) \\ p_{ij}(a,b) = \frac{1}{M} \sum_{m=1}^{M} \delta(a, s_i^m) \delta(b, s_j^m) \end{cases} \tag{2.1}$$

where p_i is the single-site frequency, that is, the normalised count of repeated amino acid a in column i of the alignment, and p_{ij} is the two-site frequency, the normalised count of the times the two amino acids a and b concurrently appear in column i and j respectively. The δ used in eq. [2.1] has the usual form:

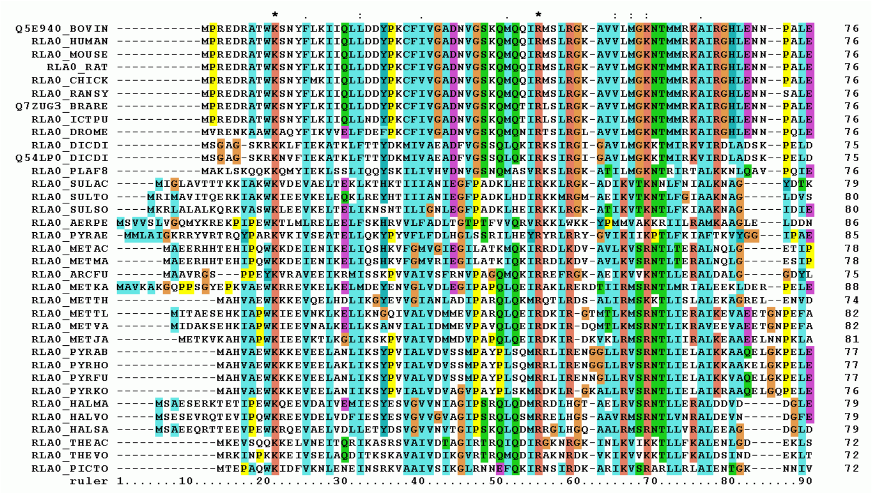


Figure 2.3: Representation of the structure of an MSA. Image taken from https://en.wikipedia.org/wiki/Multiple_sequence_alignment

$$\delta(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

An analysis of the form of these two quantities computed from the MSA can reveal important and interesting biological information about the protein family.

Considering, for example, the single site frequency p_i, if we find that this is noticeably polarised around specific values of a, that is, we can observe several amino acids of type a along the column i of the alignment, this suggests that, in the family under study, amino acid a tends to be conserved along the process of evolution the protein.

From the point of view of the biological function of the protein, this means that amino acid a in position i probably plays a key role, thanks to its chemicophysical properties, in the execution of the protein functions.

Two-site frequency p_{ij} can instead be related to the concept of coevolution of amino acid residues. Coevolution of residues is an essential process in the shaping of a protein , by understanding which residues coevolve it is possible to make better assumptions regarding a protein's shape and the functions it carries out, and even helps in identifying substitutions that may lead to desired changes in the function of the protein [22].

The kind of observations we can make by analysing the two site frequencies can regard, for example, if two amino acid residues are very often found together in two given sites, or if with the variation of one residue due to mutations, another one tends to change as well. This kind of relation hints at the possibility of the two residues being in contact in the 3D structure of the protein or generally working together to ensure the correct "behaviour" of the protein in terms of its functionality.[23] [25] So that when one residue mutates, the functionality of the protein is hindered, unless the mutation reverses or the other residue undergoes a suitable mutation as well.

We can of course practically analyse correlations among sites of the protein sequences by observing the difference between p_{ij} and $p_i p_j$, this difference, and thus the correlations, is zero if $p_{ij} = p_i p_j$, that is, if the two sites are independent.

It is important to specify, though, how even in the case of non zero correlations it cannot be said that the amino acid residues in exam certainly coevolve, as the correlations might arise from indirect influence, more refined statistical methods are used to solve this problem, distinguishing possibly coevolving residues from ones that are only linked indirectly.[24]

2.3 Potts Energy

In this section, we introduce the Potts model and its use in the course of our study.

The general idea leading to the use of the Potts model in this context, is the need for a global model able to estimate the strength of a direct contact between two residues of the protein sequence. This kind of method is generally called Direct Coupling Analysis (DCA) and it is used in several different applications such as the inference of protein residue contacts or, as in our case, the modelling of a fitness landscape. [29] [30] In the context of proteins, the Potts model is shown to outperform local measures of correlations, such as Mutual Information (MI) [27]

To determine the form of the probability of a sequence within this model, we look for the probability distribution $P(\mathbf{S})$ that maximises entropy (following the principle of maximum entropy) while reproducing the empirical observations coming from the dataset [31], in our case, those observations are the one site and two site frequencies introduced in the previous section (eq. [2.1]).

Explicitly writing the conditions we just mentioned:

$$\begin{cases} p_i(a) = \sum_{\{\mathbf{S}\}} P(\mathbf{S})\delta(a, s_i) = P_i(a) \\ p_{ij}(a,b) = \sum_{\{\mathbf{S}\}} P(\mathbf{S})\delta(a, s_i)\delta(b, s_j) = P_{i,j}(a,b) \end{cases} \tag{2.2}$$

together with equations 2.2 the normalisation condition for $P(\mathbf{S})$

$$\sum_{\{\mathbf{S}\}} P(\mathbf{S}) = 1$$

must also be imposed.

Recalling that our aim is to find $P(\mathbf{S})$ such that the entropy is maximised, we report the usual form of the Shannon entropy of the distribution:

$$S[P] = \sum_{\{\mathbf{S}\}} -P(\mathbf{S}) \log P(\mathbf{S})$$

by imposing the constraints eq. [2.2] and the normalisation condition with the use of the Lagrange multipliers formalism, we finally find the expression of the functional we need to optimise:

$$F[P] = -\sum_{\{\mathbf{S}\}} \Bigg\{ P(\mathbf{S}) \log P(\mathbf{S}) + \sum_i \lambda_i(s_i) \left[P_i(s_i) - p_i(s_i)\right] + \\ + \sum_{i<j} \lambda_{ij}(s_i, s_j) \left[P_{ij}(s_i, s_j) - p_{ij}(s_i, s_j)\right] + \Lambda \left[1 - P(\mathbf{S})\right] \Bigg\} \tag{2.3}$$

The form of the probability distribution $P(\mathbf{S})$ is now obtained as the solution of

$$\frac{\delta F[P]}{\delta P} = 0 \tag{2.4}$$

By identifying $\lambda_i(s_i) = h_i(s_i)$ and $\lambda_{ij}(s_i, s_j) = J_{ij}(s_i, s_j)$, the resulting form of P is

$$P(\mathbf{S}) = \frac{1}{Z} \exp\Bigg\{ \sum_i h_i(s_i) + \sum_{i<j} J_{ij}(s_i, s_j) \Bigg\} \tag{2.5}$$

where $Z = \sum_{\{\mathbf{S}\}} \exp\Big\{ \sum_i h_i(s_i) + \sum_{i<j} J_{ij}(s_i, s_j) \Big\}$ is the partition function ensuring the normalisation of the probability distribution.

We can interpret eq. [2.5] as an equilibrium Boltzmann distribution, consequently defining the Hamiltonian:

$$H(\mathbf{S}) = -\sum_i h_i(s_i) - \sum_{i<j} J_{ij}(s_i, s_j) \tag{2.6}$$

This energy function is the generalised Potts model. The calculations shown in this section only derive the functional form of the Hamiltonian, the actual values for the parameters h and J must be inferred from the dataset through different possible methods.

In the context of our study, assuming the values of the parameters as given, eq. [2.6] allows us to assign an energy value to a given protein sequence, providing an estimated measure of its fitness. [28] The energy landscape around a sequence, which takes into account the possible variations occurring to the protein, has in fact been shown to be informative about the effects of mutations on the protein's functionality. [26]

Chapter 3

Application of PCA method to our case of study

We now discuss the application of the Principal Components Analysis method described in Chapter 1 to our specific case of study, that is Multiple Sequence Alignments (MSA) of various protein families. After this, we will show some of the results obtained through this method.

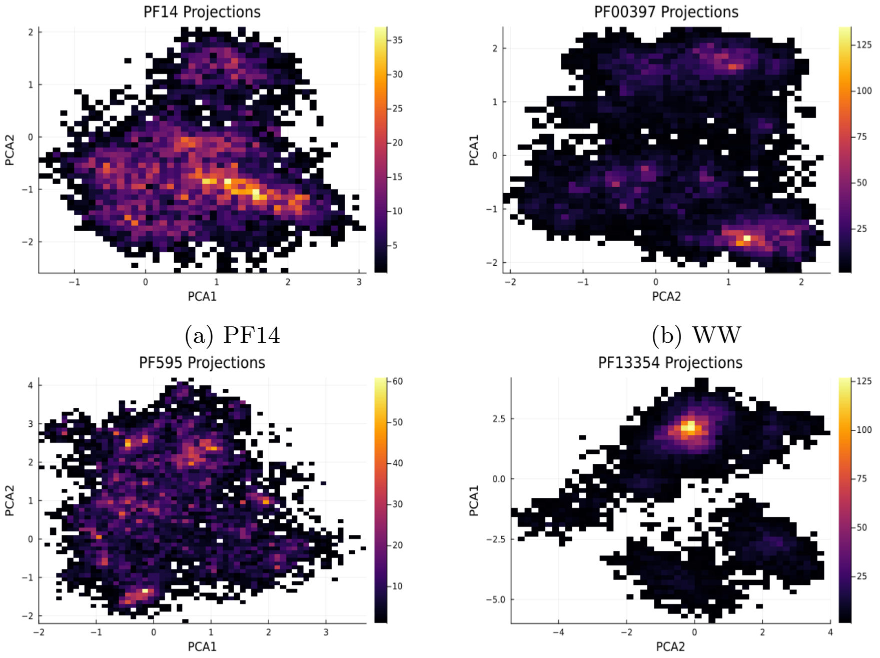
The starting dimensionality of our problem varies, as of course it is linked to the specific family we are working with. In particular, the dimensionality of the actual problem we are working with will depend on the length of the chosen protein sequence (which in our analysis is the same for all sequences in a given family) and the possible number of symbols for each position in the sequence. We will denote as L the length of a sequence in a given family and with q the number of possible symbols in each position (in our case $q = 21$). By one hot encoding the sequences we can consider both of those factors just considering the length of the one hot encoded vector. The process of one hot encoding in this case of course consists in the transformation of an L-dimensional sequence into a higher dimensional (qL) one by writing each symbol as a sequence of length q composed of only zeros and a one in the position corresponding to the value of the symbol we are considering.

Following what we said in chapter one, our aim is now to reduce the dimensionality of the problem from the starting $q \times qL$ dimensions to a number N_c of dimensions. To do so, we introduce a linear transformation of the data, where the projection matrix $\mathbf{O}$ has size $N_c \times qL$. In particular, our choice is to project the data into the first N_c components of the PCA space: therefore, the projection matrix encodes the first N_c eigenvectors of the covariance matrix computed from the MSA.

$$\mathbf{y} = \mathbf{O}\mathbf{x} \tag{3.1}$$

where $\mathbf{x} \in \{0,1\}^{q \times L}$ is the starting one hot encoded sequence and $\mathbf{y}$ is the N_c dimensional vector we wanted to obtain.

Once the N_c dimensional vector is obtained it is interesting to observe how, if $N_c = 2$ (otherwise we can consider the first two coordinates), the projections of the one hot encoded sequences in the starting natural alignment are distributed in this reduced dimensional space. We show these results for different protein families in the figure 3.1, where each subfigure depicts the density of the projections in the plane identified by the first two principal components for the various families



(a) PF14

(b) WW

(c) PF00595

(d) PF13354

Figure 3.1: Density of the projections for four different protein families MSA

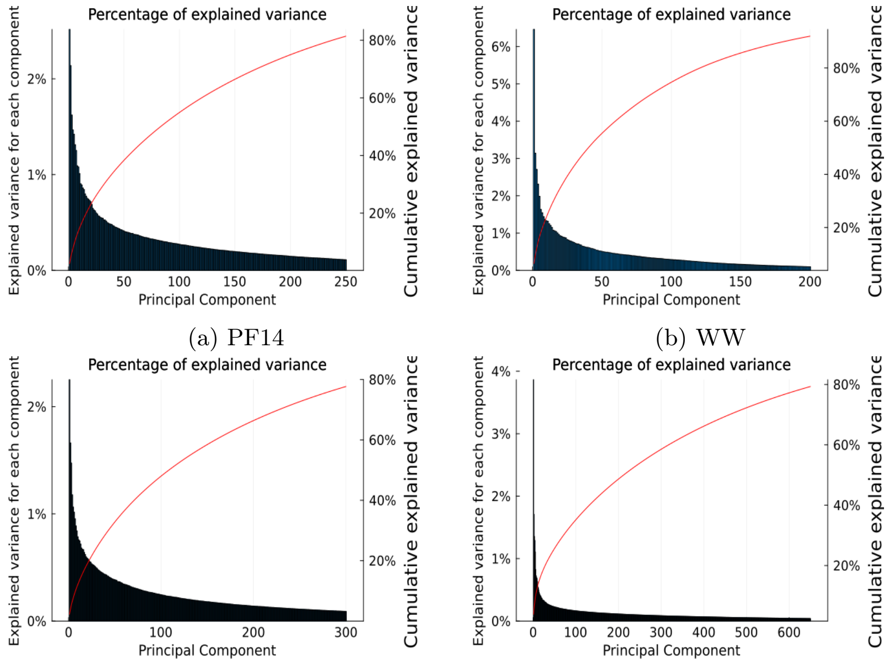
We note how these projections often form clusters, which, in the context of protein sequences, are observed to separate according to the functionality or phylogenetic relationship among sequences.

This result highlights the usefulness of dimensionality reduction, as the transformation of high-dimensional data in a low-dimensional representation often allows for a higher degree of interpretability with respect to the starting problem.

The reported images are also an example of the advantages, regarding the separation of clusters, carried by our choice of using the principal components, and their link to the variance of the dataset (see chapter 1 for a more in depth discussion).

3.1 Explained Variance

Another aspect we discussed in the theoretical introduction in chapter 1 and that we now practically present is the concept of the explained variance. The behaviour of the explained variance for the different protein families just introduced is depicted in figure 3.2



(a) PF14

(b) WW

(c) PF00595

(d) PF13354

Figure 3.2: In the graphs the percentage of explained variance associated with each principal component is shown in blue, the plot of the cumulative of the explained variance is instead depicted in red

Note that the portion of the x-axis that is shown has been chosen in such a way that the cumulative curve reaches at least 80%, the remaining components, depending on the protein family in analysis, may be numerous and contribute only negligibly to the total explained variance compared with their number.

The chosen protein families are the same as in the previous graph. In this context, it is useful to explicitly state the values of L for the different families. For PF00014 (called PF14 in the images) $L = 53$, for PF00397 (WW domain) $L = 31$, for PF00595 $L = 82$ and for PF13354 $L = 202$.

The number of principal components we are going to keep will vary depending on the protein family and on the different applications we will be focusing on, the chosen value of N_c will be specified in each different case.

By then looking at the graphs we have just shown it will then be possible to assess the percentage of explained variance in that specific case.

Chapter 4

Introduction to Expectation Propagation (EP)

Generative models as solutions of inverse problems in reduced space

Having now discussed the process through which we reduce the dimensionality of our system, the next step is the description of the analysis we will perform in the obtained lower dimensional space.

Given that, as previously mentioned, in the case of proteins the first eigenvectors can cope with a classification of the original dataset into interpretable clusters that separate according to the functionality or phylogenetic relationship among sequences. We thus assume that sequences belonging to one of the retrieved clusters have specific properties and differ from sequences of other clusters. The question we would like to ask is: Can we generate sequences revealing specific properties given only information in the reduced space?

4.1 Inverse problem in the reduced space

Recalling what we described in Chapter 1, the operation of transforming the original data in a low-dimensional representation can be formalized as a matrix product of the kind

$$\boldsymbol{y} = \mathbf{O}\boldsymbol{x} \tag{4.1}$$

where $\boldsymbol{x}$ is the qL-dimensional vector containing one of the instances of the original data, $\mathbf{O}$ is the $N_c \times qL$ rotational matrix associated with the transformation, and $\boldsymbol{y}$ is an N_c-component transformation of the original instance. In the case of the principal component analysis of protein sequences $\boldsymbol{x} \in \{0,1\}^{L \times q}$ is the one-hot encoding of a sequence of length L in an alphabet of q symbols, and the columns of $\mathbf{O}$ contain the first N_c eigenvectors of the sequence covariance matrix.

Mathematically speaking, we can rephrase the inverse problem as follows: Given a projection $\boldsymbol{y}$, namely a low-dimensional representation of an existing sequence, is it possible to generate sequences satisfying the linear relationship in Eq. (4.1)? Using Bayes theorem we formally define an a posteriori probability of the type

$$\begin{aligned} P\left(\boldsymbol{x}|\boldsymbol{y}\right) &\propto P\left(\boldsymbol{y}|\boldsymbol{x}\right)P\left(\boldsymbol{x}\right) && (4.2) \\ &\propto \mathbb{I}\left[\boldsymbol{y} = \mathbf{O}\boldsymbol{x}\right]P\left(\boldsymbol{x}\right) && (4.3) \end{aligned}$$

where $P\left(\boldsymbol{x}\right)$ is the prior probability over the target variables.

$\mathbb{I}$ introduced in eq. 4.3 is the indicator function, restricting the values of $\mathbf{y}$ to respect the projection constraints.

We first relax $\boldsymbol{x}$ to a real variable and we then set the prior probability to enforce (i) the binary nature of the original variables and (ii) the constraints carried by the one-hot encoding. Indeed, given a vector $\boldsymbol{x}$ one can retrieve a sequence only if, in each of the $q-$blocks composing $\boldsymbol{x}$, only one component is equal to one. This is equivalent to saying that the sum of the variables belonging to each of the blocks must be one. Formally, we can define a vector $\mathbf{1} = \{1\}^L$, and a matrix

$$\mathbf{A} = \begin{pmatrix} \underbrace{1,1,\ldots,1}_{q-\text{block}} & \cdots & 0 & 0 \\ 0 & \underbrace{1,1,\ldots,1}_{q-\text{block}} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \underbrace{1,1,\ldots,1}_{q-\text{block}} \end{pmatrix}$$

such that

$$\mathbf{A}\boldsymbol{x} = \mathbf{1} \tag{4.4}$$

The prior is then given by

$$P\left(\boldsymbol{x}\right) = \mathbb{I}\left[\mathbf{A}\boldsymbol{x} = \mathbf{1}\right]\prod_i \left[\rho\delta\left(x_i = 1\right) + \left(1-\rho\right)\delta\left(x_i = 0\right)\right] \tag{4.5}$$

where $\rho = 1/q$. Sampling from $P\left(\boldsymbol{x}|\boldsymbol{y}\right)$ would provide sequences satisfying the constraints in the low-dimensional space but, unfortunately, considering that in general $N_c \ll N$, the target probability is hard to determine.

4.2 Approximate solution through Expectation Propagation

We aim at approximating the probability distribution in Eq. (4.3) as a multivariate Gaussian density. Before going into the details of the approximation, we can notice that the two linear constraints in Eqs. (4.1)(4.4) can be collected in a unique system of linear equations involving the $\boldsymbol{x}-$variables and a Gaussian elimination of the matrix rows can be applied. As a consequence, we end up with two sets of variables, an independent set and a dependent set, of N_i and N_d variables respectively, linked by the linear operation

$$\boldsymbol{x}^d \quad = \quad -\mathbf{F}\boldsymbol{x}^i + \boldsymbol{z} \tag{4.6}$$

where the entries of $\mathbf{F}$ and $\boldsymbol{z}$ are the result of the Gaussian elimination operation. The exact a posteriori probability is therefore given by

$$P\left(\boldsymbol{x}^d, \boldsymbol{x}^i\right) \propto \mathbb{I}\left[\boldsymbol{x}^d = -\mathbf{F}\boldsymbol{x}^i + \boldsymbol{z}\right] \prod_{\alpha \in \{i,d\}} \prod_j \left[\rho\delta\left(x_j^\alpha = 1\right) + (1-\rho)\,\delta\left(x_j^\alpha = 0\right)\right] \tag{4.7}$$

We can then resort to a Gaussian approximation of the overall set of variables

$$Q\left(\boldsymbol{x}^d, \boldsymbol{x}^i\right) \quad \propto \quad \prod_{j=1}^{N_i} e^{-\frac{1}{2d_j^i}\left(x_j^i - a_j^i\right)^2} \prod_{j=1}^{N_d} e^{-\frac{1}{2d_j^d}\left(x_j^d - a_j^d\right)^2} \tag{4.8}$$

$$= \quad e^{-\frac{1}{2}\left[\left(\boldsymbol{x}^i - \boldsymbol{a}^i\right)^T \mathbf{D}^i \left(\boldsymbol{x}^i - \boldsymbol{a}^i\right) + \left(\boldsymbol{x}^d - \boldsymbol{a}^d\right)^T \mathbf{D}^d \left(\boldsymbol{x}^d - \boldsymbol{a}^d\right)\right]} \tag{4.9}$$

that using Eq. (4.6) becomes a function of the free variables only

$$Q\left(\boldsymbol{x}^i\right) \propto e^{-\frac{1}{2}\left[\left(\boldsymbol{x}^i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}^i - \boldsymbol{\mu}\right)\right]} \tag{4.10}$$

for

$$\begin{cases} \boldsymbol{\Sigma}^{-1} & = \mathbf{D}^i + \mathbf{F}^T \mathbf{D}^d \mathbf{F} \\ \boldsymbol{\mu} & = \boldsymbol{\Sigma}\left[\mathbf{D}^i \boldsymbol{a}^i + \mathbf{F}^T \mathbf{D}^d \left(\boldsymbol{z} - \boldsymbol{a}^d\right)\right] \end{cases} \tag{4.11}$$

The parameters of the approximation encoded in the (diagonal) matrices and vectors $\left\{\mathbf{D}^i, \mathbf{D}^d, \boldsymbol{a}^i, \boldsymbol{a}^d\right\}$ are determined by a fixed point set of equations analyzed in the following section. Note that from Eqs. (4.6)(4.10) one can determine the one-point and two-point statistics of both sets of variables as

$$\begin{array}{lclcll} \left\langle x_j^i \right\rangle_Q = & \mu_j & \left\langle x_j^{i^2} \right\rangle_Q - \left\langle x_j^i \right\rangle_Q^2 = & \Sigma_{jj} & j = & 1, \ldots, N_i \\ \left\langle x_j^d \right\rangle_Q = & \left[-\mathbf{F}\boldsymbol{\mu} + \boldsymbol{z}\right]_j & \left\langle x_j^{d^2} \right\rangle_Q - \left\langle x_j^d \right\rangle_Q^2 = & \left[\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T\right]_{jj} & j = & 1, \ldots, N_d \end{array} \tag{4.12}$$

4.2.1 Moment matching

Let us assume to be interested in the computation of the marginal probability of a specific variable, i.e. x. For the sake of simplicity, let us consider it as an independent variable. Given the full multivariate in Eq. (4.8), we can get a Gaussian marginal of the type

$$Q\left(x_j^i\right) \quad = \quad \int d^{N_d}\boldsymbol{x}^d \int d^{N_i-1}\boldsymbol{x}^i_{/x_j^i}\, Q\left(\boldsymbol{x}^d,\boldsymbol{x}^i\right) \tag{4.13}$$

As an alternative, one may consider part of the exact prior in Eq. (4.7) involving the x_j-variable: let us define the tilted distribution associated with the target variable as an approximation where we use the approximating Gaussian densities for all variables except x_j, times the exact prior enforcing the binary nature of the variable

$$\begin{aligned} Q^{(j,i)}\left(x_j^i\right) \quad \propto \quad & \left[\int d^{N_d}\boldsymbol{x}^d \int d^{N_i-1}\boldsymbol{x}^i_{/x_j^i}\, Q\left(\boldsymbol{x}^d,\boldsymbol{x}^i\right) e^{\frac{1}{2d_j^i}\left(x_j^i-a_j\right)^2}\right] \times \\ & \times \left[\rho\delta\left(x_j^i=1\right)+\left(1-\rho\right)\delta\left(x_j^i=0\right)\right] \end{aligned}$$

The integral within the parenthesis will give a univariate Gaussian density, the so-called cavity distribution $Q^{\backslash j,i}\left(x_j^i\right)$, that we parametrize by a mean $\mu^{\backslash j,i}$ and a variance $\Sigma^{\backslash j,i}$. Let us consider now the general case of a variable x_j^α for $\alpha=\{i,d\}$; the tilted distribution and the univariate Gaussian can be computed respectively as

$$Q^{(j,\alpha)}\left(x_j^\alpha\right) \propto e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2}\left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right] \tag{4.14}$$

$$Q\left(x_j^\alpha\right) \propto e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} e^{-\frac{1}{2d_j^\alpha}\left(x_j^\alpha-a_j^\alpha\right)^2} \tag{4.15}$$

Since part of the exact prior is encoded in the tilted distribution, we may assume that this approximation will be more accurate than Eq. (4.13); this observation can be exploited to tune the parameters $\left(a_j^\alpha, d_j^\alpha\right)$. Ideally, we aim at getting the best set of parameters such that Eq. (4.15) is as close as possible to the tilted in Eq. (4.14); formally we can minimize the Kullback-Leibler distance between the two. [32] The explicit computation is reported in Appendix A. The result of the minimization is the moment matching condition:

$$\begin{aligned} \left\langle x_j^\alpha \right\rangle_Q \quad &= \quad \left\langle x_j^\alpha \right\rangle_{Q^{(j,\alpha)}} \\ \left\langle {x_j^\alpha}^2 \right\rangle_Q \quad &= \quad \left\langle {x_j^\alpha}^2 \right\rangle_{Q^{(j,\alpha)}} \end{aligned}$$

Since this holds for all variables, we get a set of fixed point equations that can be used to iteratively update the unknown parameters of the approximation $\left\{\mathbf{D}^i,\mathbf{D}^d,\boldsymbol{a}^i,\boldsymbol{a}^d\right\}$

$$\begin{cases} d_j^\alpha & = \left(\frac{1}{\left\langle {x_j^\alpha}^2 \right\rangle_{Q^{(j,\alpha)}} - \left\langle x_j^\alpha \right\rangle_{Q^{(j,\alpha)}}^2} - \frac{1}{\Sigma^{\setminus j,\alpha}} \right)^{-1} \\ a_j^\alpha & = d_j^\alpha \left[\left\langle x_j^\alpha \right\rangle_{Q^{(j,\alpha)}} \left(\frac{1}{d_j^\alpha} + \frac{1}{\Sigma^{\setminus j,\alpha}} \right) - \frac{\mu^{\setminus j,\alpha}}{\Sigma^{\setminus j,\alpha}} \right] \end{cases} \tag{4.16}$$

This is a very general result; the details of the problem only enter in the type of prior we are considering and, therefore, in the computation of the first and second moments of the tilted distributions. Note that, in principle, one has to compute $N_i + N_d$ Gaussian integrals to get the parameters of the cavity $\left(\mu^{\setminus j,\alpha}, \Sigma^{\setminus j,\alpha}\right)$ for all $j's$ of the two sets, but this costly operation can be skipped by noting that the cavity parameters are linked to the parameters of the full Gaussian distribution in Eq. (4.10) by

$$\begin{cases} \Sigma^{\setminus j,i} = & \frac{\Sigma_{jj}}{1-\Sigma_{jj}\left(d_j^i\right)^{-1}} \\ \mu^{\setminus j,i} = & \frac{\mu_j - a_j^i \Sigma_{jj}\left(d_j^i\right)^{-1}}{1-\Sigma_{jj}\left(d_j^i\right)^{-1}} \end{cases} \qquad \begin{cases} \Sigma^{\setminus j,d} = & \frac{\left[\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T\right]_{jj}}{1-\left[\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}\right]_{jj}\left(d_j^d\right)^{-1}} \\ \mu^{\setminus j,d} = & \frac{\left[-\mathbf{F}\boldsymbol{\mu}+\boldsymbol{z}\right]_j - a_j^d \left[\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T\right]_{jj}\left(d_j^d\right)^{-1}}{1-\left[\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T\right]_{jj}\left(d_j^d\right)^{-1}} \end{cases} \tag{4.17}$$

4.3 Generative EP

Given a particular instance, i.e. a natural sequence $\boldsymbol{x}^{(n)}$, we can devise an Expectation Propagation approximation of $P\left(\boldsymbol{x}|\boldsymbol{y}^{(n)}\right)$. At convergence, we can use the approximate Gaussian in Eq. (4.10) to sample the independent variables and then, using Eq. (4.6), we can obtain the associated dependent variables. Let us call the final result as the sample $\boldsymbol{x}^{(t)}$ composed of

$$\begin{aligned} \boldsymbol{x}^{(t),i} & \sim \mathcal{N}_{(n)}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ \boldsymbol{x}^{(t),d} & = -\mathbf{F}\boldsymbol{x}^{(t),i} + \boldsymbol{z} \end{aligned}$$

The quality of the sampled sequences depends on (i) the number of components N_c used to obtain the projections and, possibly, (ii) the “temperature” at which the sampling is performed. Indeed, one can use

$$\begin{aligned} \boldsymbol{x}^{(t),i} & \sim \mathcal{N}_{(n)}\left(\boldsymbol{\mu}, T\boldsymbol{\Sigma}\right) \\ \boldsymbol{x}^{(t),d} & = -\mathbf{F}\boldsymbol{x}^{(t),i} + \boldsymbol{z} \end{aligned}$$

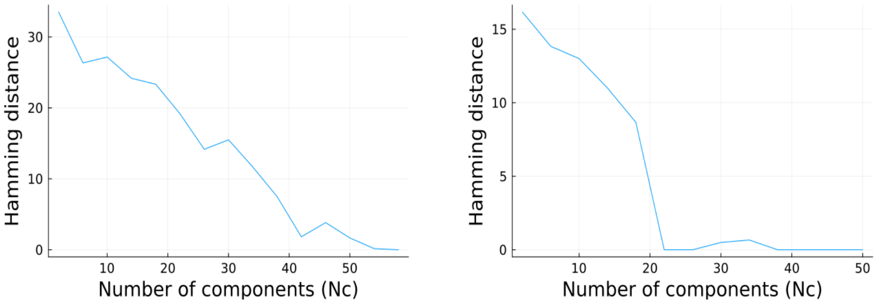
for $T \neq 1$.

Chapter 5

Application of the Expectation Propagation (EP) method

In this chapter we will discuss the results obtained via the implementation of the method described in chapter 4. In particular the scope of this section is only to evaluate how well the method is able to generate qL dimensional sequences starting from N_c dimensional projections, while tweaking the controllable parameters of the method. Leaving the actual analysis of the protein families and the distribution of the sequences in the lower dimensional space to the next chapters.

We start by considering the Hamming distance between the starting natural sequence and the sequence obtained from the application of the Expectation Propagation method on the N_c dimensional projection of the starting sequence. This is done for different values of N_c and the result is shown in figure 5.1



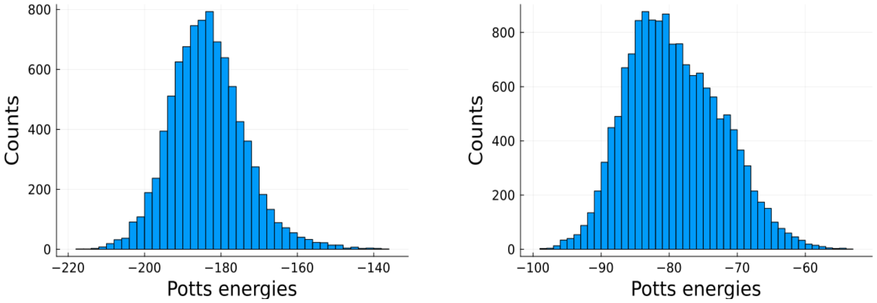
(a) PF14

(b) WW

Figure 5.1: Hamming distances of the sampled sequences compared to the starting natural sequences

The results, as expected, show that for an increasing N_c the Hamming distance decreases, as the "reconstruction" of the original sequence improves. For some value of N_c the Hamming distance clearly goes to zero, marking the threshold over which we observe a perfect reconstruction of the original sequence. The graphs just shown depict an average of the behaviour of 6 randomly picked sequences. We now focus on two of the protein families illustrated in Chapter 3,in particular, we will now focus on PF14 and PF397 (WW domain)

To give some context before the results shown in the next graphs, we first illustrate the distribution of the energy of the natural sequences for both PF14 and WW alignment:

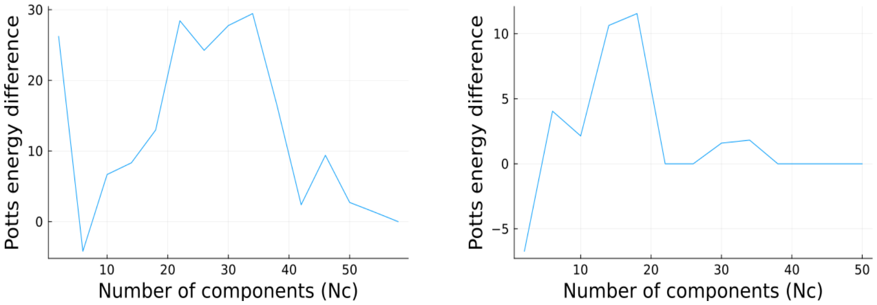


(a) PF14

(b) WW

Figure 5.2: Distribution of Potts energies of the natural sequences for PF14 and WW

We now show graph similar to the one shown in figure 5.1, regarding instead the difference in terms of Potts energy between the sequences obtained via the application of EP and the natural sequences, results are illustrated in figure 5.3



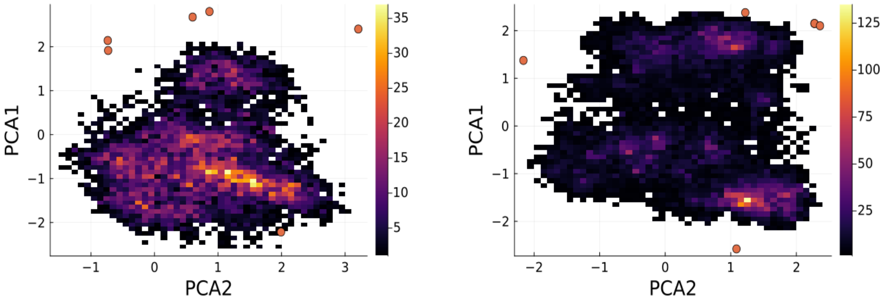
(a) PF14

(b) WW

Figure 5.3: Energy difference between the sampled sequences and the starting natural sequences

The results are coherent with what we already observed in Figure 5.1 regarding the Hamming distance. In this case, the energy of the sequence obtained with the application of EP tends to the energy of the starting natural sequence, as in the case of the previous graph we obtain a perfect reproduction for N_c larger than some threshold that we will call N_c^*, this value clearly varies depending on the protein family we are considering and will be important in the next chapters to identify the point in which we are operating along the working range of the method.

We also notice how, in the plots just shown, energy values for a small number of N_c are lower than the energy of the starting natural sequence. This is an interesting recurring behaviour, as the obtained sequences, despite the good results in terms of Potts energies, actually correspond to projections in the two dimensional PCA space which are not close to the projections of the starting natural sequences and are not distributed in a way which seems coherent with the clusters formed by the projections of the sequences in the MSA under analysis, as shown in figure 5.4.



(a) PF14

(b) WW

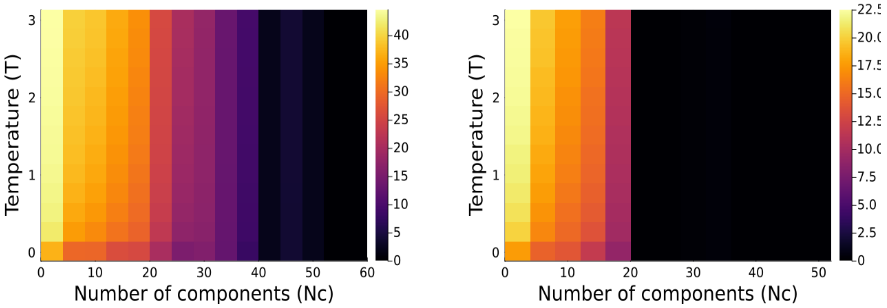
Figure 5.4: The projections of the obtained sequences with better energy than the natural sequences shown in figure 5.3

The points of the scatter plot shown in the figure represent the projections on the N_c-dimensional space of the sequences obtained from the application of the EP method on the projections of the same natural sequences used to compute the plot in figure 5.3. In particular, since we are interested in the projections of the obtained sequences that display a lower Potts energy with respect to the starting natural proteins, we chose sequences obtained for $N_c = 6$ in the case of PF14 and $N_c = 2$ for WW. Those values correspond, in fact, to the portion of the plots in figure 5.3 in which the energy difference is smaller than 0.

We can observe how those projections are indeed positioned out of the area occupied by the clusters of projections of the natural sequences of the alignment.

5.1 Application of generative EP

Much in the same way we can study the results obtained through the generative EP approach described at the end of chapter 4. In this case, another variable apart from N_c can be modified, that is the "temperature" T. In figure 5.5 we show once again the difference in Hamming distance between natural and sampled sequences for different protein families

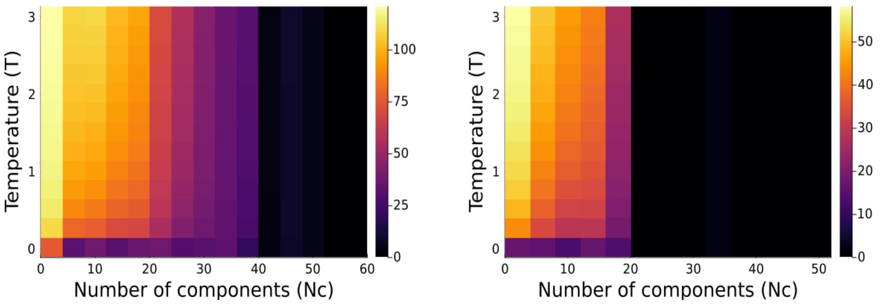


(a) PF14

(b) WW

Figure 5.5: Hamming distances of the sampled sequences with respect to the starting natural sequences, the colour shown in the graph represents the Hamming distance

and the difference in terms of Potts energies is shown in figure 5.6 . As in the case of the previous graphs we represented the energy of the sampled sequences minus the energy of the starting natural sequences, so that a negative value means that the energy of the sampled sequence is lower with respect to the natural sequence.



(a) PF14

(b) WW

Figure 5.6: Energy difference between the sampled sequences and the starting natural sequences, the colour shown in the graph represents this Potts energy difference

In figures 5.3 and 5.4 the existence of a critical value of N_c, identifying the threshold over which we obtain a perfect reconstruction of the original sequence starting from its projection in the reduced dimensional space, is even clearer than in the graphs shown in the previous paragraph. This value of N_c is generally close to the value of L and doesn't seem to vary for different values of T.

The influence of T becomes clearer in the range of N_c smaller than N_c^*, with an higher temperature causing a worsening of the obtained results.

Chapter 6

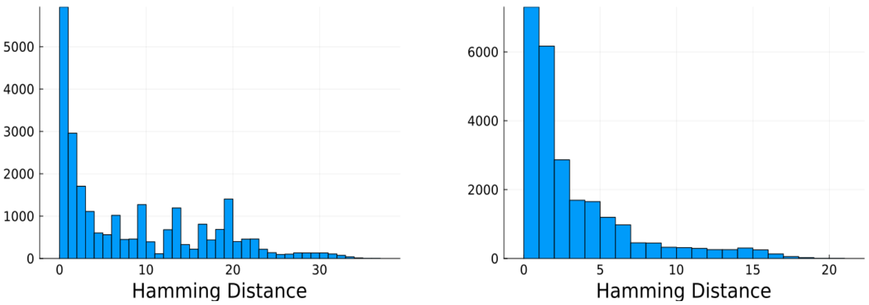
Results for several protein families

We now show some of the results obtained with the application of the previously described methods to some of the protein families introduced in chapter 3 where we described the results obtained through the application of PCA.

As we saw in the chapter about the application of the EP method to our case, the quality of the reconstruction of the starting sequence and its consistency largely depends on the parameters T and (especially) on the number of dimensions of the reduced-dimensional space on which we projected the sequences in the context of Principal Components Analysis, that is what we called N_c. Depending on which point of the working range of EP application we showed in chapter 6 we decide to operate in, different results can be obtained. Firstly we focus on the results for values of N_c smaller than N_c^*, where N_c^* is the critical valuefor N_c we introduced in chapter 5, in this case the application of the method is mainly sampling, starting from the natural sequences in the original alignment we project them onto the N_c-dimensional space as described in chapter 3, compute, using EP, the probability distribution of the original sequence and use it to generate a given number of sampled sequences. In the following we shall present the main features of the results obtained in this context.

6.1 Analysis of the statistics of the sampled sequences

To start we show a more complete depiction of the actual distribution of the variables considered in the graphs regarding the working range for the EP algorithm, as in that case only partial information in conveyed . For example, the distribution of the Hamming distance between the sampled sequences and the starting (natural) sequence is shown in figure 6.1



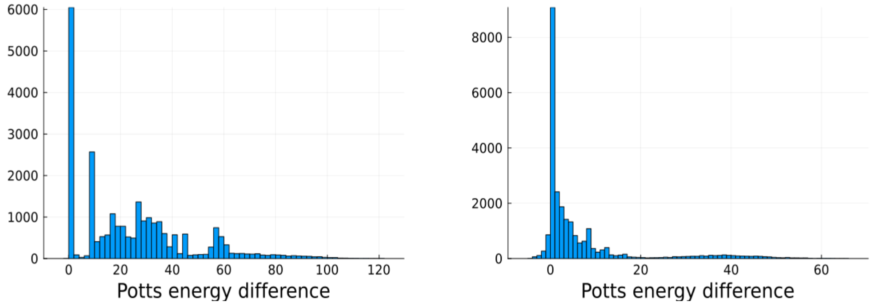
(a) PF14

(b) WW

Figure 6.1: Hamming distances of the sampled sequences with respect to the starting natural sequences

each of those graphs again referring to PF14 and WW families was done considering 50 starting sequences and computing 500 samples using the distributions obtained via the EP algorithm at N_c equal to 40 for PF14 and 31 for WW domain, with T equal to 2 for both families

Similarly, we show the graph regarding the difference in terms of Potts energies between the sampled sequences and the starting (natural) sequence



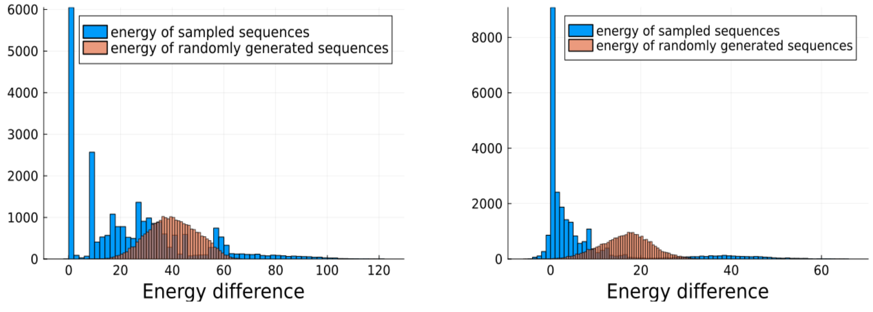
(a) PF14

(b) WW

Figure 6.2: Potts energy difference between the sampled sequences and the starting natural sequences

All of the graphs in figures 6.1 and 6.2 are clearly peaked around 0, representing the case in which we manage to sample exactly the same sequence whose projection we started from. We will briefly return to these results in the next section about walks, as at that point the reason why the graphs take this form will be clearer.

A first way to test the quality of these results is comparing the histograms shown in figures 6.1 and 6.2, with histograms obtained in the same way but that are instead computed using randomly generated sequences of comparable Hamming distance from the natural sequences. This is depicted in figure 6.3



(a) PF14

(b) WW

Figure 6.3: Comparison of the energy differences for the sampled sequences and of the randomly generated sequences

it is clear how the distribution of the energy is much better (the average energy is lower) in the case of the sequences generated via the application of PCA and EP with subsequent sampling than in the case of randomly generated sequences. This tells us how the method can retain the important biological information about the protein sequence (so it keeps a low Potts energy) when "passing" through a reduction of dimensionality as imposed by PCA.

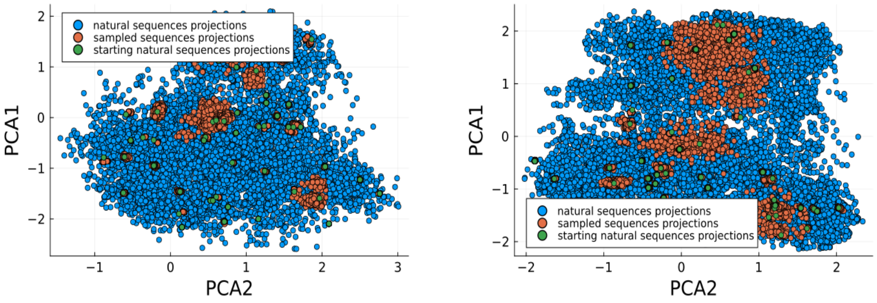
6.1.1 Distribution of samples' projections in the first two coordinates of the reduced dimensional space

Given the aim of the application of PCA and EP in this regime, that is, sampling, it is interesting to analyse the projections of the sequences obtained from said sampling, on the N_c dimensional space, comparing it with the original N_c-space distribution of the projections of the natural sequences. It is evident how some of these projections cannot be represented fully, particularly the ones for which $N_c > 3$. In those cases we will represent the first two coordinates, that is the ones that exhibit the clusterization in the projection of the natural sequences shown in chapter 4.

These results are illustrated in figure 6.4.

the graphs were done in the same conditions as the previous ones, starting from the same 50 natural sequences and computing 500 samples at the same values of N_c and T. We see how the samples' projections are mainly spread around their respective starting sequence, in the sampling we thus retain the information regarding position in N_c-dimensional space.

As before, we now compare these results with the same graphs obtained using random sequences of comparable Hamming distance with respect to the

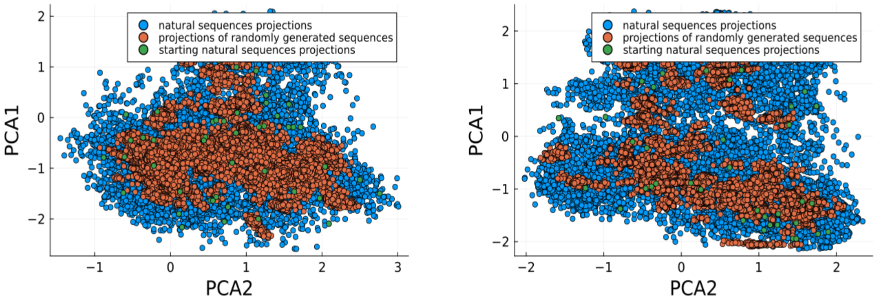


(a) PF14

(b) WW

Figure 6.4: Comparison of projections of the sampled sequences and of the natural sequences

starting sequence. This is shown in figure 6.5



(a) PF14

(b) WW

Figure 6.5: Comparison of projections of the randomly generated sequences and of the natural sequences

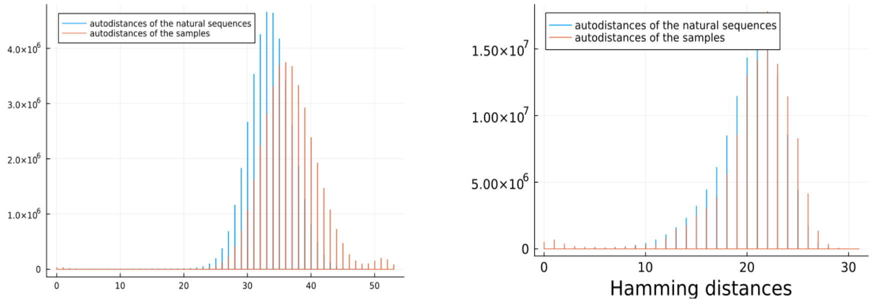
It is clear how in this case all information regarding the position of the starting sequence is lost.

Another notable feature of the sampled alignment is the autodistance of its sequences, that is the Hamming distances between each of the sampled sequences, and its comparison with the autodistance in the starting (natural) alignment.

We can visualise the autodistances of an alignment as a matrix $\mathbf{D}$ of dimensions $M \times M$ in which the entry on the i-th row and j-th column is equal to the Hamming distance between the i-th and the j-th sequences of the MSA, that is, the i-th and j-th rows of the matrix representing the alignment. All of the entries of $\mathbf{D}$ (except for the ones on the diagonal) are then plotted as an histogram.

We expect to reproduce a similar behaviour for the autodistances of the two alignments.

The two behaviours again for PF14 and WW domain are represented in figure 6.6, which are simple histograms of the computed autodistances representing explicitly each value of the autodistance for both the natural and sampled alignment.



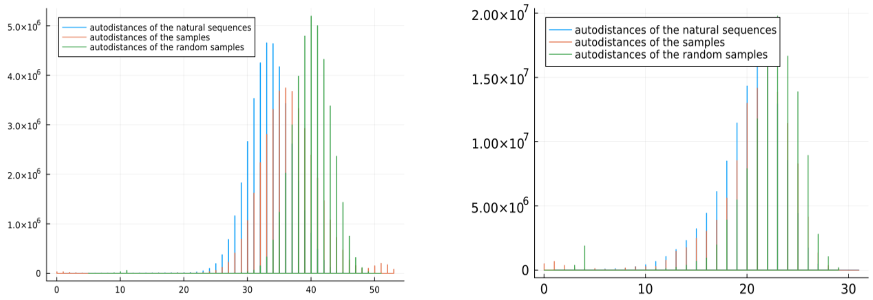
(a) PF14

(b) WW

Figure 6.6: Autodistances of the original alignment and of the sampled alignment

The obtained distribution of the autodistances in the case of the sampled alignment is similar to the one obtained from the natural sequences, in this case, the results were obtained starting from 200 sequences and sampling 100 sequences for each of them at values of N_c equal to 40 for PF14 and 31 for WW.

As we did in the case of the variation of energy, also in this context a comparison with the autodistance computed on an alignment of randomly generated sequences can be useful to verify the quality of the obtained results. This is shown in figure 6.7



(a) PF14

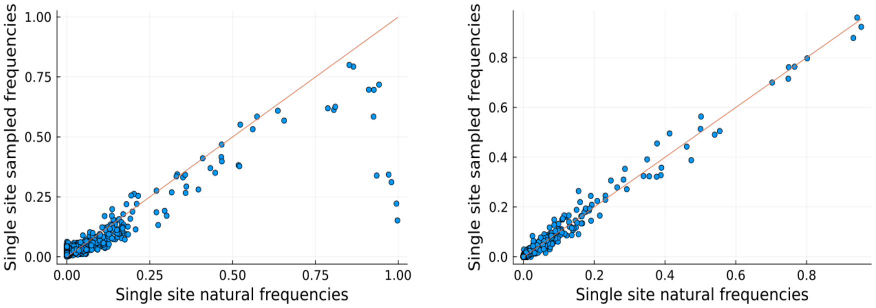
(b) WW

Figure 6.7: Autodistances of the original natural alignment, the sampled alignment and of the random one

6.1.2 Single site and two site frequencies

An important check, regarding the statistics of the obtained sampled sequences, is the comparison of single site frequencies and two site frequencies between the original (natural) alignment and the ones we computed. In particular, to take into account the behaviour of the two site frequencies we show the matrix **C**, whose entries are defined as $c_{ij} = p_{ij} - p_i p_j$. Of course the result we would want to obtain is a certain degree of correspondance between those two quantities, to highlight this, we represent each of the components of the frequencies in a scatter plot, in which we plot on the x-axis the natural frequencies and on the y-axis the ones computed from the sampled sequences.

It is clear how, given that the order of both vectors containing the frequencies is the same, the more the two vectors are similar the more the points in the plot will be placed along the diagonal direction. We show those graphs in figure 6.8



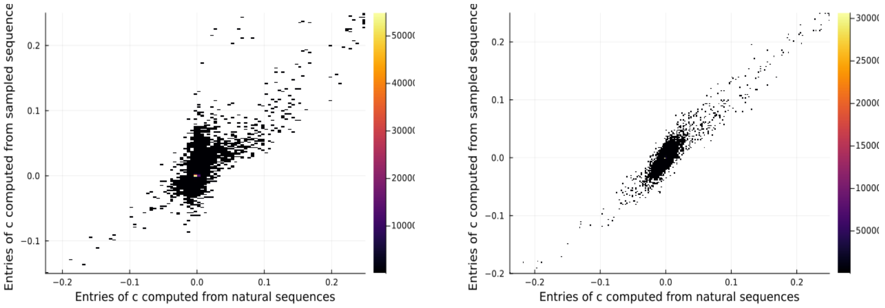
(a) PF14

(b) WW

Figure 6.8: Comparison of the single site frequencies computed from the natural alignment and from the sampled alignment

The comparison between entries of the matrix **C** is instead shown in figure 6.9 Note how, to plot the two site frequencies in the same way as the single site ones, the matrix containing the data has to be "flattened" to a vector, because of the high number of points we also chose to show the results using a density plot instead of the scatter plot as in figure 6.8.

The results for WW family behave as expected, with the points of the two plots generally aligned with the plotted diagonal line, indicating a good reproduction of the single and two site statistics in the sampled alignment. In the case of PF14 the method seems to fail in reproducing the correct statistical properties, we in fact obtain a much higher single site frequency for some residues that are not that present in the original MSA and a very noisy plot regarding the entries of **C**.



(a) PF14

(b) WW

Figure 6.9: Comparison of the two site frequencies computed from the natural alignment and from the sampled alignment

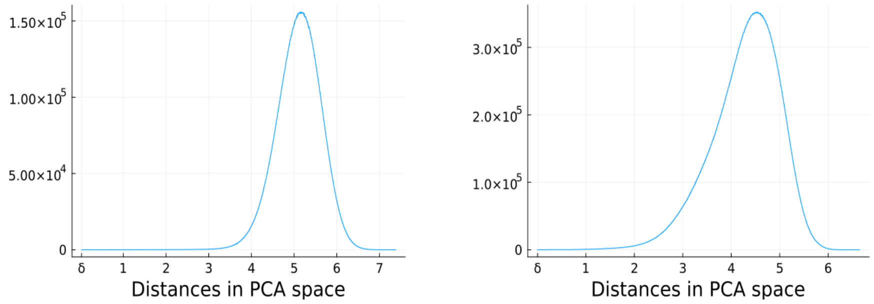
Chapter 7

Random walks on N_C-dimensional space

After talking, in chapter 6, about the results obtained in the range characterized by $N_c < N_c^*$, in this chapter we will focus on the possible results in the case $N_c > N_c^*$. In particular we saw how in this regime the relation between the projected natural sequences and the natural sequences in the full dimensional space is bijective, we can use this result to explore the N_c-dimensional space in which those projections live.

We start by briefly commenting on how the direction and length of the step are chosen, starting from the latter. The first thing we need to specify is the unit of length used to measure the displacement. To determine this, all of the Euclidean distances (in the N_c-dimensional space) between each pair of sequences in the natural alignment are computed and the minimum among them is chosen as the unit of length that we call δ, which will be used to describe all displacements in this section. Notice how the actual value of δ is different for each protein family. In each analysis the appropriate value of the unit length will of course be used. Results for the distribution of distances between all sequences in the natural alignment and the value of δ are shown in figure 7.1 for PF14 and WW domain families.

It is interesting to observe how the large majority of the sequences are at a distance of approximately 5 (depending on the protein family) from each other, the tail of the distribution arrives to values of distance much smaller than the peak (this minimum value is what we identified as δ) but includes only a very small portion of the total number of proteins in the alignment. Another fundamental aspect that needs to be highlighted is that the actual resulting modulus of the displacement,in the context of this study, is dependent on the direction , so, when talking about the choice of the length of the step, the parameter that can be controlled is actually the isotropic displacement (measured in terms of δ), which will then be weighted by the anisotropic contribution coming from the directionality and by an isotropic random contribution.



(a) PF14

(b) WW

Figure 7.1: Euclidean distances between all sequences in the natural alignment

We now explain the origin of those two added contributions, the random element is, quite simply, a Normally distributed random variable weighing each component of the vector representing the displacement (a different random variable is used for every component), the second one is instead linked to the information provided by the reduced space landscape, each component of the displacement vector is multiplied by the square root of the corresponding eigenvalue of the covariance matrix obtained from the alignment.

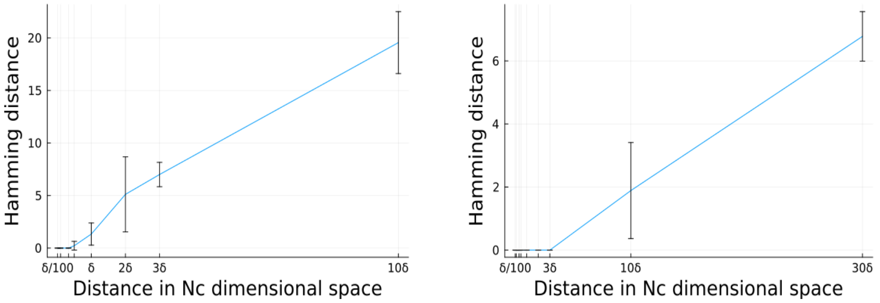
To sum it up, the displacement is computed as: $x_i^{t+1} = x_i^t + \sqrt{\lambda_i} * W * n\delta$ with $i = 1, 2, ..., N_c$

where x_i^t is the i-th component of the coordinate at step t, λ_i is i-th eigenvalue of the covariance matrix, W is the Normal distributed random variable and n is the length of the chosen isotropic displacement measured in "times of δ"

7.1 Variation of Hamming distance with respect to a displacement in N_c-dimensional space

A first interesting result to observe in the analysis of the movement in the lower dimensional space is the relation between the distance between two points in the N_c-dimensional space and the resulting difference, in terms of Hamming distance, between the sequences (in the full dimensional space) corresponding to those points. In particular, this is done starting from projections of sequences from the natural alignment, generating a step with an arbitrary length and along an arbitrary direction (in agreement with what we said in the previous paragraph) following which we move to a different point in the N_c-dimensional space, then a sequence is obtained via the EP algorithm starting from the coordinates of the point we moved to and the Hamming distance is computed for this sequence with respect to the natural sequence we started from. This procedure is repeated several times for each sequence and then the same is done for different sequences.

Graphs resulting from this kind of analysis are shown in figure 7.2



(a) PF14

(b) WW

Figure 7.2: Hamming distance with respect to PCA distance

The plots shown in Figure 7.2 show the results obtained by considering several starting sequences, and considering then various different steps starting from those sequences. We notice how the behaviour clearly varies among the two different families, with the value of δ obtained for both PF14 and WW domain still having a meaningful role in determining the distance over which the algorithm is still able to reconstruct the original sequence.

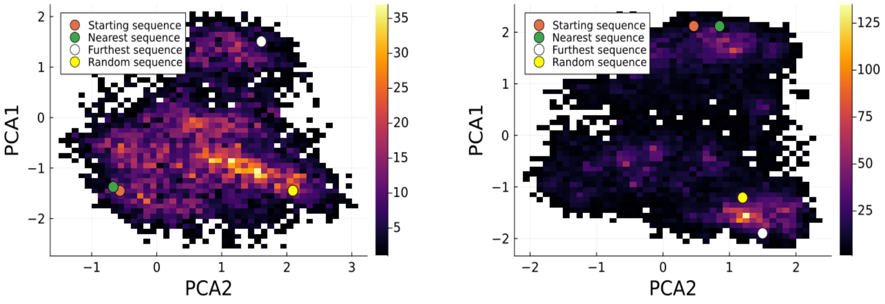
7.2 Walks along a line connecting two natural sequences

Once the space surrounding the sequences has been analyzed in terms of Hamming distances, we can now switch to the analysis of the Potts energy landscape in the N_c-dimensional space. As a first method, we consider the study of a walk going in a straight line from one known (natural) sequence to another known sequence, each of the points in this walk is then studied applying the EP algorithm and extracting the wanted information (in this case the Potts energy) from the obtained sequence.

Notice how, in order to correctly generate the straight path connecting the two chosen sequences, in this case we will need to eliminate the anisotropic contribution to the step generation that was put in through the eigenvalues of the covariance matrix, as previously explained. In this context we will then use a value of a step such that each point in the discrete walk is equally spaced.

In order to give a more complete view of the typical landscape of the Potts energy in the reduced dimensional space we will consider, for each sequence, three different walks. The first is the walk between the chosen sequence and the natural sequence that is closer to it (in the N_c-dimensional space), the second is the walk between the chosen sequence and the furthest one, and finally the walk between the chosen sequence and a random one from the alignment.

To make it easier to visualise the process, we show in figure 7.3 the first two coordinates of the projections of the chosen natural sequences, highlighting the starting point and the three different target sequences. These points are plotted over the density graph representing the projections of all the sequences of the Multiple Sequence Alignment, as the ones showed in chapter 3. We also specify the values of Euclidean distance in the N_c-dimensional space and the Hamming distance between the starting sequence and the three target ones. For PF14 the Euclidean distance between the chosen starting sequence and the nearest sequence is 0.86 with Hamming distance 4, the distance from the furthest sequence is 6.41 with Hamming distance 40 , while for the randomly picked target sequence we find a distance of approximately 5.4 and an Hamming distance equal to 35. For WW, only reporting the values following the same order as in the case of PF14, 1.18 and 6, 5.94 and 23, 5.2 and 23.



(a) PF14

(b) WW

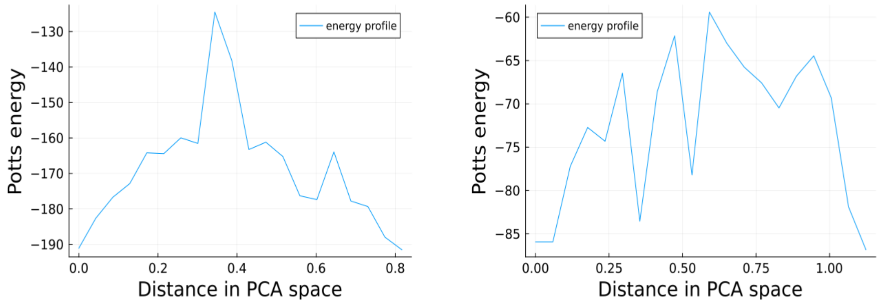
Figure 7.3: First two coordinates of the projections of the natural sequences involved in the walk described in section 7.2

We start by showing the case of the walk to the nearest sequence, some examples of the typical behaviour of the Potts energy of the reconstructed sequences along the line connecting the two sequences are shown in the graphs below (figure 7.4). The direction of the movement is from left to right of the graph.

Graphs in figure 7.5 now illustrate the case of the walk to the furthest sequence from the chosen starting point, and finally in figure 7.6 we show the case of the walk to a randomly picked sequence.

As evident from these graphs, the kind of Potts energy landscape which seems to be resulting from this analysis is composed of "peaks" of energy located around the middle of the line connecting the two sequences, and wells of varying width depending on the specific sequence, having most of the time a minimum corresponding to the coordinate of the target sequence itself.

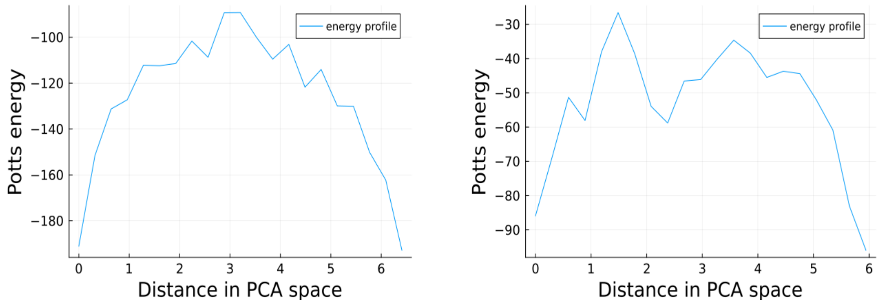
The difference between the three analysed cases is also noticeable, with energy peaks along the walk being considerably higher in the case of sequences



(a) PF14

(b) WW

Figure 7.4: Potts energy behaviour along the line connecting the two natural sequences, this is the case of a randomly picked sequence and the one nearest to it



(a) PF14

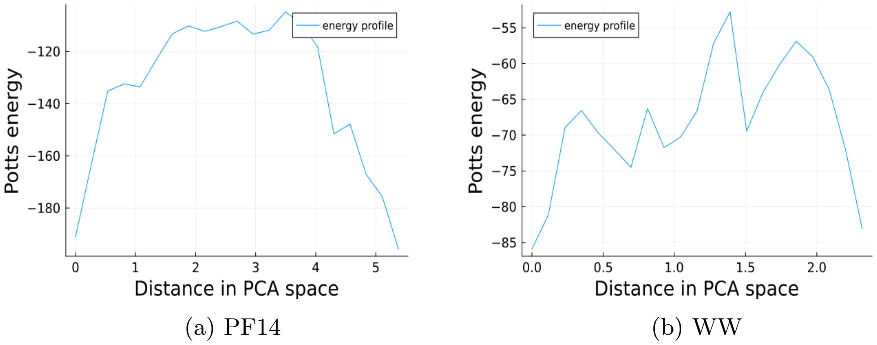
(b) WW

Figure 7.5: Potts energy behaviour along the line connecting the two natural sequences, this is the case of a randomly picked sequence and the one furthest from it

which are more distant in the reduced dimensional space.

At this point we can go back to the results obtained in the previous section about sampling, and give an interpretation. What we observed in the graphs regarding the energy difference with respect to the starting (natural) sequence (Figure 6.2) was a distribution peaked around small energy difference values, with energies of the samples being almost always higher than the one of the natural sequence energy. Here we found how this result is coherent with the characteristics of the energy landscape we just analysed, that is, given the constraint of a local sampling around the starting sequence and a structure of the energy landscape formed by wells positioned in correspondence to the natural sequences, a distribution in the form of the one we obtained is to be expected.

It is also interesting to show the graphs regarding the behaviour of the Hamming distance of the sequences along the walk with respect to both the

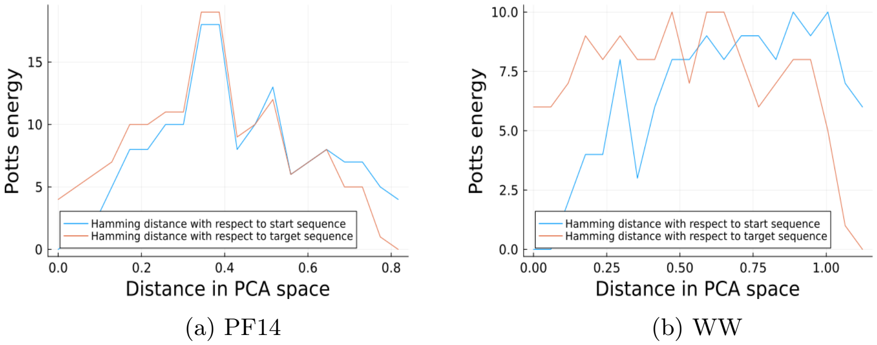


(a) PF14

(b) WW

Figure 7.6: Potts energy behaviour along the line connecting the two natural sequences, in this case both of the sequences are picked randomly

starting and the target sequence. This is shown in figure 7.7.



(a) PF14

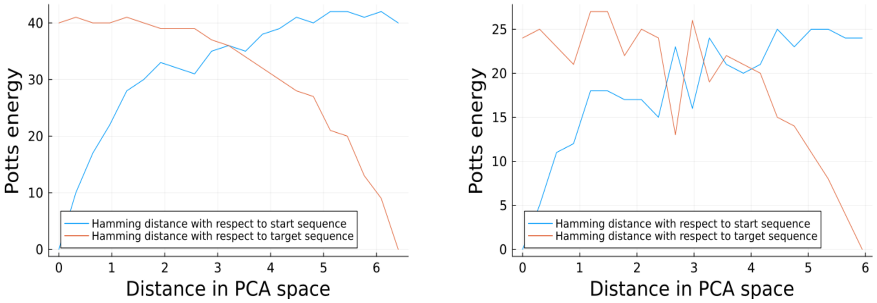
(b) WW

Figure 7.7: Hamming distance behaviour along the line connecting the two natural sequences, this is the case of a randomly picked sequence and the one nearest to it

These graphs are computed for the same sequences as the previous ones. The result is of course as expected, with one of the curves starting from the value of the Hamming distance between the two natural sequences and going to zero, and the other displaying the exact opposite behaviour. Also in this case the range of variation of the Hamming distance along the walk considerably varies in three cases we considered, though this could be due to the initial difference in Hamming distance among the three selected sequences and the starting one.

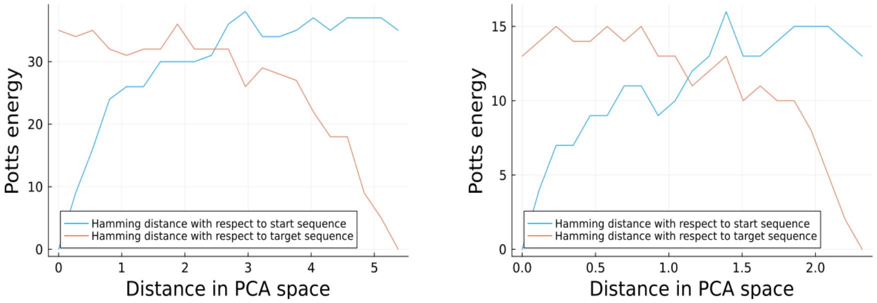
7.2.1 Effect of the variation of N_c on the Potts energy landscape

The results shown in this section seem to indicate a structure of the energy landscape constituted of wells in correspondence of a natural protein sequence



(a) PF14 (b) WW

Figure 7.8: Hamming distance behaviour along the line connecting the two natural sequences, this is the case of a randomly picked sequence and the one furthest from it



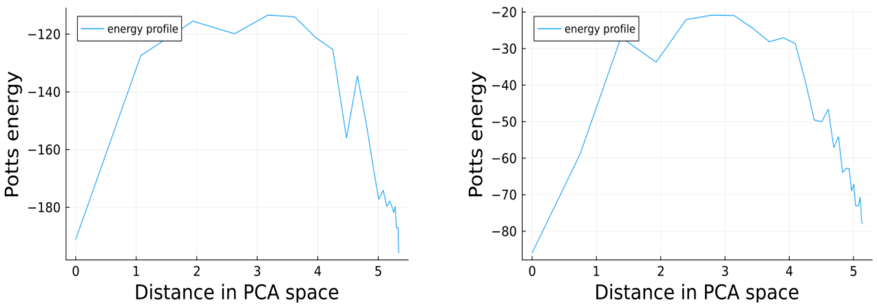
(a) PF14 (b) WW

Figure 7.9: Hamming distance behaviour along the line connecting the two natural sequences, in this case both of the sequences are picked randomly

and higher energy peaks in the space between them. In order to better understand the structure of the landscape near these points, and later its dependence on the number of components N_c, it is interesting to perform the same analysis we just showed with an increased number of steps in the neighborhood of the target sequence. These results are reported in figure 7.10

The same study is then performed with an increased value of N_c, in particular, we chose $N_c = 250$ for PF14 and $N_c = 200$ for WW. The graphs in figure 7.11 depict the results found in this case.

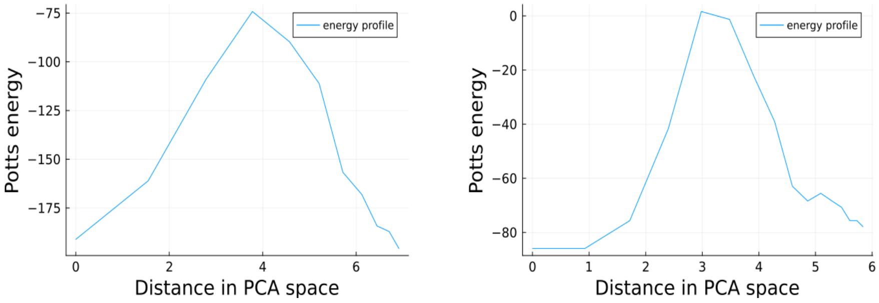
Aside from observing that the value of the distance between the two natural sequences is different, due to the variation in the dimensionality of the space, the most important information we can deduce from this graph is that a higher value of N_c effectively widens the range inside which EP converges to the natural sequences, we also observe an increase in the amplitude of the energy well located in correspondence of the natural sequence under analysis.



(a) PF14

(b) WW

Figure 7.10: Energy profile along the line connecting the two chosen natural sequences



(a) PF14

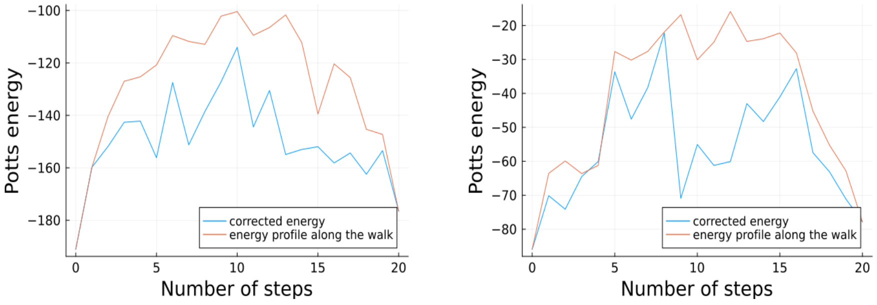
(b) WW

Figure 7.11: Energy profile along the line connecting the two chosen natural sequences, we have now increased the number of components N_c

7.3 Montecarlo method for optimisation of the energy

The final analysis we perform regarding the study of the Potts energy landscape in the reduced dimensional space is the use of a Monte Carlo random walk, in an attempt to optimise the energy of a given protein sequence by moving its associated projection in the N_c-dimensional space.

In particular, we will apply this method to the same walk analysed in the previous section, our aim is to optimise each of the points of the walk, in this way we attempt to find the path of minimal energy between two given natural sequences. In the plot 7.12 we show the energy profile obtained in the same way as the one in figure 7.6 and the new energy profile resulting from the optimisation process. Notice how both the profiles are plotted with respect to the number of the steps, while in section 7.2, we showed the behaviour along the line connecting the two sequences, this is of course because the new steps of the walk are no longer positioned along the same line as the old steps as a consequence of the optimisation.

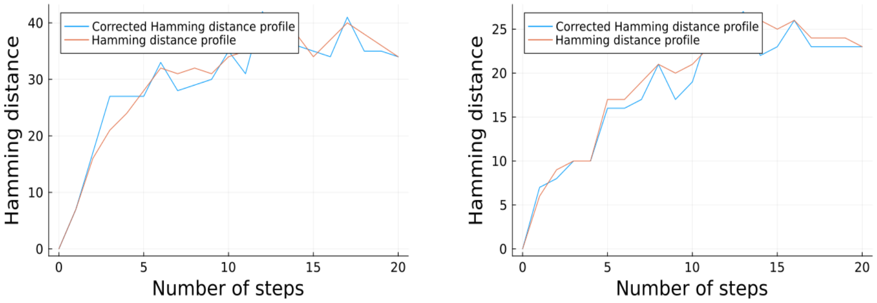


(a) PF14

(b) WW

Figure 7.12: Energy profiles of the optimised and unoptimised walks

We can see how the Monte Carlo random walk is able to visit sequences with generally lower energy along the path. For completeness we also show the variations in the profile of the Hamming distance (see figure 7.13)



(a) PF14

(b) WW

Figure 7.13: Hamming distance profiles of the optimised and unoptimised walks

Interestingly, the Hamming distance of the optimised sequences is, for almost all the steps, very close to the value of the unoptimised ones.

An increase in the number of steps of the walk would be beneficial in order to prove the existence of a minimum energy path connecting the two sequences.

Some interesting observations we can make about the shown results, though, are that it is, in general, possible to lower the Potts energy starting from a given point in the reduced dimensional space, and that as a result of the optimisation we found sequences with Potts energies comparable with the energies of the proteins belonging to the MSA.

Chapter 8

Conclusions

In this thesis we presented, on chapters 1 to 4, the methods, the characteristics of our data and the structure of the dataset. In chapters 5 and 6 we discussed the properties of the mapping between the full dimensional space and the N_c dimensional one, observing, for example, the variations in the results of the application of EP for different values of N_c, or the statistics of alignments of sequences sampled starting from the projections of the natural sequences in the reduced dimensional space.

By observing these results we concluded that, over a certain value of N_c, dependent on the specific protein family, the mapping between the two spaces is bijective. The value of N_c can be then chosen in accord with the application of the method we are interested in. The sampling process is of course also dependent on the chosen value of N_c.

Once we analysed the properties of the mapping, we focused on the analysis of the reduced dimensional space, computing the effect (in terms of Hamming distance) of a displacement in the N_c dimensional space or studying the Potts energy landscape in which the proteins live. We recall how the Potts energy was chosen to represent a measure of the fitness of a given sequence, observations on such energy landscape assume then particular relevance.

In this regard, we found how the energy landscape in the reduced dimensional space seems to be characterised by wells of low energy, positioned in correspondence of the projections of the natural sequences, while higher energy sequences are found in the space between the wells.

Some of the possible further results in a future perspective include a refinement of the application of the Monte Carlo method, which we basically just proved to work in general terms. An analysis of the predicted protein contacts performed on the sampled sequences could be an interesting test of the predictive power of this method.

Appendix A

Appendix A

We report here the explicit derivation of the moment matching condition shown in chapter 4, our aim is to minimize the Kullback-Leibler divergence, which in this case takes the form:

$$
\begin{aligned}
&D_{KL}(Q^{(j,a)}(x_j^\alpha)||Q(x_j^\alpha)) = \\
&= \int dx_j^\alpha \frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right] \times \\
&\times \log(\frac{\frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right]}{\frac{1}{Z_Q} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} e^{-\frac{1}{2d_j^\alpha}\left(x_j^\alpha-a_j^\alpha\right)^2}})
\end{aligned}
\tag{A.1}
$$

Taking the derivative with respect to the parameter a_j^α:

$$
\begin{aligned}
&\frac{\partial D_{KL}(Q^{(j,a)}(x_j^\alpha)||Q(x_j^\alpha))}{\partial a_j^\alpha} = \\
&= -\int dx_j^\alpha \frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right] \times \\
&\times \frac{\frac{\partial}{\partial a_j^\alpha}\frac{1}{Z_Q} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} e^{-\frac{1}{2d_j^\alpha}\left(x_j^\alpha-a_j^\alpha\right)^2}}{\frac{1}{Z_Q} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} e^{-\frac{1}{2d_j^\alpha}\left(x_j^\alpha-a_j^\alpha\right)^2}} = \\
&= -\int dx_j^\alpha \frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right] \frac{1}{d_j^\alpha}(x_j^\alpha-a_j^\alpha) + \\
&+ \int dx_j^\alpha \frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\backslash j,\alpha}}\left(x_j^\alpha-\mu^{\backslash j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha=1\right)+\left(1-\rho\right)\delta\left(x_j^\alpha=0\right)\right] \frac{Z_Q'}{Z_Q}
\end{aligned}
\tag{A.2}
$$

By putting the form of the derivative just found equal to 0 we find:

$$\frac{1}{d_j^\alpha}\langle x_j^\alpha\rangle_{Q^{(j,\alpha)}} - \frac{1}{d_j^\alpha}a_j^\alpha =$$

$$= \int dx_j^\alpha \frac{1}{Z_{Q^{(j,\alpha)}}} e^{-\frac{1}{2\Sigma^{\setminus j,\alpha}}\left(x_j^\alpha - \mu^{\setminus j,\alpha}\right)^2} \left[\rho\delta\left(x_j^\alpha = 1\right) + (1-\rho)\,\delta\left(x_j^\alpha = 0\right)\right] \left(\frac{1}{d_j^\alpha}\langle x_j^\alpha\rangle_Q - \frac{1}{d_j^\alpha}a_j^\alpha\right)$$

$$\frac{1}{d_j^\alpha}\langle x_j^\alpha\rangle_{Q^{(j,\alpha)}} = \frac{1}{d_j^\alpha}\langle x_j^\alpha\rangle_Q \tag{A.3}$$

The same calculation, with very similar steps, can be done for the derivative with respect to the parameter d_j^α, here we only report the final result:

$$\frac{\partial D_{KL}(Q^{(j,a)}(x_j^\alpha)||Q(x_j^\alpha))}{\partial d_j^\alpha} = 0$$

$$\frac{1}{2(d_j^\alpha)^2}\langle (x_j^\alpha)^2\rangle_{Q^{(j,\alpha)}} - \frac{1}{(d_j^\alpha)^2}a_j^\alpha\langle x_j^\alpha\rangle_{Q^{(j,\alpha)}} = \frac{1}{2(d_j^\alpha)^2}\langle (x_j^\alpha)^2\rangle_Q - \frac{1}{(d_j^\alpha)^2}a_j^\alpha\langle x_j^\alpha\rangle_Q \tag{A.4}$$

Combining the two obtained results (eq. A.3 and A.4) we clearly find the moment matching condition

$$\begin{aligned} \left\langle x_j^\alpha \right\rangle_Q &= \left\langle x_j^\alpha \right\rangle_{Q^{(j,\alpha)}} \\ \left\langle x_j^{\alpha^2} \right\rangle_Q &= \left\langle x_j^{\alpha^2} \right\rangle_{Q^{(j,\alpha)}} \end{aligned}$$

Bibliography

[1] Ian T. Jolliffe and Jorge Cadima *Principal component analysis: a review and recent developments*, The Royal Society publishing. (2016) https://doi.org/10.1098/rsta.2015.0202

[2] Rasmus Bro a and Age K. Smilde *Principal component analysis*, The Royal Society of chemistry. (2014) 10.1039/C3AY41907J

[3] Svante Wold, Kim Esbensen, Paul Geladi *Principal component analysis*, Elsevier. (1987) https://doi.org/10.1016/0169-7439(87)80084-9

[4] Johnstone IM, Lu AY. *On Consistency and Sparsity for Principal Components Analysis in High Dimensions.* J Am Stat Assoc. 2009 Jun 1;104(486):682-693. doi: 10.1198/jasa.2009.0121. PMID: 20617121; PMCID: PMC2898454.

[5] Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space.* https://doi.org/10.1080/14786440109462720

[6] I. T. Jolliffe *Principal Component Analysis.* Springer New York, NY. https://doi.org/10.1007/b98835

[7] Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, 24(6), 417-441. https://doi.org/10.1037/h0071325

[8] Horn RA, Johnson CR. *Hermitian and symmetric matrices. In: Matrix Analysis.* Cambridge University Press; 1985:167-256.

[9] Sanvictores T, Farci F. *Biochemistry, Primary Protein Structure.* [Updated 2022 Oct 31]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK564343/

[10] PAULING L, COREY RB. *Atomic coordinates and structure factors for two helical configurations of polypeptide chains.* Proc Natl Acad Sci U S A. 1951 May;37(5):235-40. doi: 10.1073/pnas.37.5.235. PMID: 14834145; PMCID: PMC1063348.

[11] Walter Kauzmann. *Structural factors in protein denaturation.* (May 1956) Journal of cellular Physiology https://doi.org/10.1002/jcp.1030470410

[12] Walter Kauzmann. *Some factors in the interpretation of protein denaturation.* (1959) Advances in Protein Chemistry Volume 14. 10.1016/S0065-3233(08)60608-7

[13] Milner-White EJ. *Protein three-dimensional structures at the origin of life. Interface Focus.* 2019 Dec 6;9(6):20190057. doi: 10.1098/rsfs.2019.0057. Epub 2019 Oct 18. PMID: 31641431; PMCID: PMC6802138.

[14] Van Holde, K. E., Mathews, Christopher K. *Biochemistry* (1996) https://archive.org/details/biochemistry00math

[15] Murray RF, Harper HW, Granner DK, Mayes PA, Rodwell VW. *Harper's Illustrated Biochemistry.* (2006) New York: Lange Medical Books/McGraw-Hill.

[16] Nelson DL, Cox MM. *Principles of Biochemistry (4th ed.).* (2005) New York: W. H. Freeman

[17] Carrillo, H., Lipman, D. *The Multiple Sequence Alignment Problem in Biology.* (1988) https://doi.org/10.1137/0148063

[18] Nuin PA, Wang Z, Tillier ER. *The accuracy of several multiple sequence alignment programs for proteins.* BMC Bioinformatics. 2006 Oct 24;7:471. doi: 10.1186/1471-2105-7-471. PMID: 17062146; PMCID: PMC1633746.

[19] Szalkowski AM. *Fast and robust multiple sequence alignment with phylogeny-aware gap placement.* BMC Bioinformatics. 2012 Jun 13;13:129. doi: 10.1186/1471-2105-13-129. PMID: 22694311; PMCID: PMC3495709.

[20] Xiang Z. *Advances in homology protein structure modeling.* Curr Protein Pept Sci. 2006 Jun;7(3):217-27. doi: 10.2174/138920306777452312. PMID: 16787261; PMCID: PMC1839925.

[21] Koonin, E. V. (2005). *Orthologs, Paralogs, and Evolutionary Genomics 1.* (2005). https://doi.org/10.1146/annurev.genet.39.073003.114725

[22] Wang, M., Kapralov, M.V., Anisimova, M. *Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco.* BMC Evol Biol 11, 266 (2011). https://doi.org/10.1186/1471-2148-11-266

[23] Ju, F., Zhu, J., Shao, B. et al. *CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction.* Nat Commun 12, 2535 (2021). https://doi.org/10.1038/s41467-021-22869-8

[24] Alexander Fung, Antoine Koehl, Milind Jagota, Yun S. Song bioRxiv 2022.10.16.512436; doi: https://doi.org/10.1101/2022.10.16.512436

[25] Little DY, Chen L. *Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination in Protein Evolution.* (2009) PLoS ONE 4(3): e4762. https://doi.org/10.1371/journal.pone.0004762

[26] Rodriguez Horta, E.; Barrat-Charlaix, P.; Weigt, M. *Toward Inferring Potts Models for Phylogenetically Correlated Sequence Data.* Entropy 2019, 21, 1090. https://doi.org/10.3390/e21111090

[27] Eleonora De Leonardis, Benjamin Lutz, Sebastian Ratz, Simona Cocco, Remi Monasson, Alexander Schug, Martin Weigt, *Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction*, Nucleic Acids Research, Volume 43, Issue 21, 2 December 2015, Pages 10444-10455, https://doi.org/10.1093/nar/gkv932

[28] Richard R. Stein ,Debora S. Marks,Chris Sander . *Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models* (2015) https://doi.org/10.1371/journal.pcbi.1004182

[29] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A. 2011 Dec 6;108(49):E1293-301. doi: 10.1073/pnas.1111471108. Epub 2011 Nov 21. PMID: 22106262; PMCID: PMC3241805.

[30] Ivan Anishchenko, Petras J. Kundrotas, and Ilya A. Vakser *Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model* (2018) https://doi.org/10.1016/j.bpj.2018.07.035

[31] Shimagaki, Kai and Weigt, Martin *Selection of sequence motifs and generative Hopfield-Potts models for protein families* , Phys. Rev. E , American Physical Society, Volume 100, Issue 3, September 2019, doi:10.1103/PhysRevE.100.032128

[32] Minka P. Thomas, *Expectation Propagation for Approximate Bayesian Inference*, (2013) arXiv:1301.2294