LUND
UNIVERSITY

Testing Direct Coupling Analysis on HP Model Proteins

Author: Mustafa Kadhim **Supervisor:** Anders Irbäck

Department of Astronomy and Theoretical Physics

January 18, 2018

Abstract

Direct coupling analysis (DCA) models correlations in sets of related (homologous) protein sequences using a Potts-like spin model ansatz. From the couplings of the Potts model, derived by inverse statistical mechanics, residue-pair contacts in the 3D structure of the protein are predicted. In this thesis, this approach is applied to structures from the HP model on a square lattice. All HP sequences folding to the structures studied are known from previous work. For the calculation of the couplings, a maximum likelihood procedure is implemented, based on gradient descent and Monte Carlo methods.

Popular Science Summary

In decades, we have been trying to understand the DNA, RNA and the proteins responsible for several functions in our bodies and organisms. Functions such as, reproduction of cells, transmitting information and transporting of molecules.

Since the DNA and RNA chains can be quite complex and long, we find it easier to deal with protein chains. Also, we know that proteins can help in binding and forming the DNA helix chains, thus it is sufficient to understand the behaviour of proteins and what provide their functions. After further research into the subject, it has been discovered that a protein's structure play an important role in deciding its function and where to operate inside a body.

A typical protein chain consists of a large number of organic molecules, amino acids, sitting next to each other, as a necklace. Each amino acid is interacting with other amino acids in the chain and sometimes with other nearby protein chains, depending on the circumstances. These types of interactions can develop into a complicated mathematical problem, that needs to take many variables into account. Also, the complexity of the calculations grows exponentially with the number of amino acids in the chain. Scientists have been trying to solve the problem of predicting a protein's structure by knowing its amino acid components. Many methods have been tried and proposed. One of these methods require taking advantage of inverse statistical procedures such as a direct coupling analysis (DCA) methodology, where it is considered useful to reveal contacts between amino acids in the protein sequence, in which later on can be used to predict the structure of the protein. The DCA method compares many protein sequences with each other, aiming to find a correlation between the amino acids that influence the protein's appearance (shape). From this method, it turns out that one often can infer pairs of amino acids in contact. Knowledge of such contacts can greatly facilitate structure prediction.

If we succeed to predict the structure of proteins, only by knowing their amino acid sequences, we can start designing our own proteins, able to function as we desire at the required location in the body. Furthermore, we will be able to understand more about cancer, tumours and neurodegenerative diseases.

In conclusion, determining protein shapes can have a huge beneficial impact on our life, from increasing our life expectancy, to answering fundamental questions in biology. Thus, it is an important problem that we need to solve.

Contents

Acknowledgement

Big and sincere thanks to prof. Irbäck for the supervising, encouragement and being always available to guide me through this thesis. I am also grateful to Jeremy and Dolev, for making the time to support me when I needed something. Finally, thank you very much Harry, for the inspiration, being always there to discuss my ideas and teaching me how to express them in a better way.

1 Introduction

Imagine a teacher having a class with a certain number of students. The teacher suggests a game to play. As a preparation, the students are asked to pick a number being either 1 or -1. Then, they are requested to form a line, placing one hand on the shoulder of the student in front of them and keeping their second hand free. This game has two rules: firstly, students are not allowed to step off the line to interact with their friend standing in a different location, not next to them in the line. The only way to interact with that person is by forcing the whole line to fold itself so that an interaction, such as shaking hands can occur. Secondly, no interactions are allowed between students having the same position index type, that is, no interaction between a student having an even position index and other students with an even index, and also, no interaction between a student with an odd position index and others having an odd index.

When the game begins, the students start to change the structure of the line while having these rules in mind. After a while, the line has been re-formed to a shape, allowing the students to interact with their friends who at the same time satisfy the second rule. As observed from the game, the stronger the friendship of the students, the harder they try to alter the shape of the line, as they attempt to shake hands with their friends.

This game mimics a toy model for one of the unsolved problems in modern science, the *protein folding problem*. Solving it will allow scientists to predict the *native tertiary structure* of the proteins by knowing their *primary structure*. Proteins can be thought of as the line of students. Every student in the line represents an amino acid, while the chosen number by that student represents the amino acid type. In reality, there are 20 naturally occurring amino acids from which proteins are built [1]. The sequence of 1 and -1 along the line is what we define as the primary structure. Later, when the game starts and the students have reshaped the line, this shape corresponds to the so-called tertiary structure of the protein. For many proteins, the preferred (native) tertiary structure is the state of minimum Gibbs free energy, which is the state that determines the function of proteins under physiological conditions [2]. Storing and transferring molecules, accelerating chemical reactions, transporting information between cells and operating as structural building blocks, are all biological activities carried out by proteins [3]. In general, proteins are divided into three main types, namely membrane, fibrous and globular proteins. Unlike membrane and fibrous proteins, globular proteins are water soluble. In aqueous solution, they tend fold into compact (globular) structures (their native states) [4].

For the protein chains to fold, a force must be involved. In fact, there are several non-covalent forces acting upon the protein to drive folding and preserve its native structure; for instance, hydrogen bonding, charge-charge interactions and van der Waals interactions [5]. In addition, the largely entropy-based hydrophobic effect is known to play a key role [6, 7]. Some of the amino acids building up proteins are apolar or hydrophobic (interact only weakly with water), while other ones are polar or hydrophilic (interact favorably with water). As a result, hydrophobic residues tend to cluster together in the interior of globular

proteins, while polar residues tend to dominate on the surface of the protein.

In the students game analogy, the students chose labels 1 or -1. Suppose a given pair of students want to form a contact if and only if, both carry the label 1. Students with this label will then tend to cluster together and end up in the core of the structure, much like the hydrophobic residues of a protein in water.

A challenge that remains, is the ability to determine the native 3D (tertiary) structure. Methods such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) are used to give detailed structure analyses [8]. However, these methods are both time consuming and expensive. A long-standing goal, therefore, is to develop viable computational approaches for predicting the 3D structure of proteins from their linear sequences of amino acids [4], often referred to as *protein structure prediction* (PSP).

PSP can be roughly divided into two different problems, homology modeling and free modeling. Homology modeling means that it is possible to find a sequence which is related (homologous) to the given one and whose structure is known. This structure can then serve as a starting point, which greatly simplifies the structure prediction problem. Free modeling refers to cases in which no such related sequence can be found, and is much more difficult.

A recent and important step forward in the area of free modeling is the development of the *direct coupling analysis* (DCA) methodology [9, 10]. DCA uses evolutionary information to predict residue-pair contacts in protein structures, without assuming access to any known structure. The method gives factual predictions about finding residue pairs having explicit mutual restrictions while evolving, and therefore are likely to be in close contact in the 3D structure. It has been found that the contacts predicted by DCA can be very useful in PSP.

In this thesis, we implement and apply DCA-like methods to study sequences from a toy model for proteins, the lattice-based so-called HP model introduced by Lau and Dill [11]. For short chains in this model, all sequences folding to a given structure are known [12].

2 The HP Model

To gain insight into the basics of protein folding, toy models such as the HP model can be very useful. The HP model is a lattice-based model, in which protein chains are represented as self-avoiding strings of beads on a lattice [13, 14]. In our study, a 2D square lattice is used. Each bead represents one amino acid and can be of two types, either hydrophobic (H) or polar (P). Multiple beads are not allowed to occupy the same position on the lattice. The use of a two-letter alphabet is a drastic simplification, and one could argue that not all amino acids can be categorized simply as hydrophobic or polar. Despite this, it turns out that there exist HP sequences that are protein-like in the sense that they possess a unique structure (a unique ground state).

The HP model tends to favour hydrophobic interactions and consider them to be the leading cause behind folding. This is based on the observation that hydrophobic residues cluster inside the molecules while enclosed by polar residues [5]. Nevertheless, hydrophobic interactions minimizes the free energy of a protein, which is a direct consequence of letting their cooperation have a negative energy value.

Communication between the amino acids in the lattice, takes the form of a non-adjacent contact, whereas no interaction occurs among residues located as neighbours in the sequence. Two non-adjacent amino acids interact through being spatially contiguous in the structure (contact) as shown in figure 1.

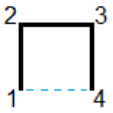

Figure 1: Illustrates a contact (blue line) between amino acids 1 and 4 on a 2D-square lattice.

An additional constraint due to the geometry of the lattice is parity. The parity constraint means that no amino acids with an odd index interacts with other amino acid residues having an odd index. The same principle holds for residues having an even index. This is illustrated in detail by figure 2.

Figure 2: The parity approved interactions between the residues, including a contact (blue line) between amino acids 0 and 3 on a 2D-square lattice.

In comparison with the teacher game, it is the students disliking water who contributes the most to the folding of the line into a specific structure, so they can both shake hand with their friends (sharing a similar dislike of water) and avoid exposing to water by gathering in the middle, surrounded by students enjoying interacting with water. When it comes to the constraints, students are not allowed to occupy the same spot in the line since each student should have a unique position index, and no hand shaking with the neighbour fellow student nor with other students having the same position index parity (even or odd).

3 Direct Coupling Analysis (DCA)

By cause of advancement in the genomic sequencing methods, more genetic data became available. As a result, more than 10^7 protein sequences were discovered [15]. These proteins can be restricted to specific groups, where they share the same structure, function and possibly the same progenitor. Such groups are described as *protein families*. Within the families, homologous sequences (linked through evolution) can be found. These sequences are considered to be a useful source of statistical information, playing a significant part in solving the *Protein Structure Prediction* (PSP) problem.

Due to gaps and insertions, homologous protein sequences are not perfectly aligned with each other to emphasize affinity in the same family [16]. To overcome this obstacle, a sequence alignment method such as *Multiple Sequence Alignment*(MSA) is used to take insertions and gaps into consideration. However, in our case the extracted sequences from the HP database do not suffer from gaps or insertions, thus there is no need to apply MSA. The technique of MSA is aligning the postulated to be homologous sequences in a list, then, arranging and matching them by sorting sequences in rows and examine the amino acids in each column. Its purpose is ascertaining the sequences to be evolutionarily related, and finding out the conserved positions and sequence correlations. These correlations demand that residues co-evolve through time to preserve the protein structure. Any alteration in conserved locations could change the proteins shape.

Unfortunately, discovering the correlations turned out to be unhelpful in unravelling the PSP complexity, as correlations might arise from direct or intermediate interactions between the amino acids in the sequence chain. A method for distinguishing between these correlations was needed. Fortunately, the method of *Direct Coupling Analysis* (DCA) can in part solve this issue. To reproduce measured sequence correlations from MSA, DCA uses a Potts model ansatz. The Potts model is given by

$$P(\sigma_1, .., \sigma_N) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1 \leq i < j \leq N} J_{ij}(\sigma_i, \sigma_j) \right) \tag{1}$$

where $h_i(\sigma_i)$ stand for the tendency of an amino acid σ_i to be found at position i, $J_{ij}(\sigma_i, \sigma_j)$ denotes the interaction between amino acids (σ_i, σ_j) at positions i, j and Z is the normalization constant. The parameters $\{h_i, J_{ij}\}$ can later on be evaluated to match the measured data. In the students game, the parameter $\{h_i\}$ indicates the tendency of a student at i, to choose a σ_i as its number, and $\{J_{ij}\}$ reveals the strength of friendship between students at i and j, having σ_i and σ_j as their chosen numbers.

This requirement of evaluating the parameters in (1), relates it to *Inverse Statistical Mechanics* (ISM), since it exploits observations to modify its parameters, aiming to regenerate experimental data, instead of predicting the outcome of experiments.

The data produced by DCA could be used as a guide when structural aspects of proteins are being examined, as it highlights the couplings that dominates in the protein. Being

aware of them can assist in predicting the protein structure[16].

4 Method

4.1 Applying DCA to HP proteins

Given a protein family of B aligned sequences, the needed empirical data for the Potts model can be extracted as follows; let $\boldsymbol{\sigma}=(\sigma_1, \sigma_2, .., \sigma_N)$ be a sequence of length N, where $\sigma_i \in \{1, .., 21\}$ is an integer indicating the amino acid type. The interval of σ_i results from nature having 20 known amino acids that most protein consist of, plus, the possibility of having a gap between the residues. For the HP model proteins, $\sigma_i = \{\pm 1\}$, considering only the hydrophobic (H) and hydrophilic (P) characteristics. The following values are taken by σ_i depending on the amino acid flavour,

$$\sigma_i = \begin{cases} +1 & \text{if H} \\ -1 & \text{if P} \end{cases}$$

which reduces the Potts model to an Ising model.

The sequences from the database are aligned so that every sequence occupy a certain row and every amino acid is aligned with other amino acids at the same sequence position. Due to this alignment, the empirical data for **the average of single** σ_i and **the correlations between** σ_i **and** σ_j are calculated from the input sequences as

$$\langle \sigma_i \rangle^B = \frac{1}{B} \sum_{b=1}^{B} \sigma_i^{(b)} \tag{2}$$

$$\langle \sigma_i \sigma_j \rangle^B = \frac{1}{B} \sum_{b=1}^{B} \sigma_i^{(b)} \sigma_j^{(b)} \tag{3}$$

where $\sigma_i^{(b)}$ denotes the amino acid at sequence position i in input sequence b.

At conserved sequence positions, $\langle \sigma_i \rangle^B$ takes the values

$$\langle \sigma_i \rangle^B = \begin{cases} +1 & \text{only H} \\ -1 & \text{only P} \end{cases}$$

To decide the parameters $\{J_{ij}, h_i\}$, a *maximum likelihood* method is introduced. Its purpose is to maximize the probability of observing the B input sequences in the Potts model. To do so, we require a negative log-likelihood function to be minimized. The negative log-likelihood function is given by

$$L = -\frac{1}{B}\sum_{b=1}^{B} \log \mathbf{P}(\sigma^{(b)}) \tag{4}$$

and equivalent to

$$L = \log Z - \sum_{i<j} J_{ij}\langle \sigma_i \sigma_j \rangle^B - \sum_i h_i \langle \sigma_i \rangle^B. \tag{5}$$

Minimizing eq. (5) implies finding the critical points satisfying the conditions

$$\frac{\partial L}{\partial h_i} = \langle \sigma_i \rangle - \langle \sigma_i \rangle^B = 0 \tag{6}$$

$$\frac{\partial L}{\partial J_{ij}} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \sigma_j \rangle^B = 0 \tag{7}$$

where $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$ are averages in the model defined by (1), that is

$$\langle \sigma_i \rangle = \sum_{\sigma} \sigma_i \mathbf{P}(\sigma) \tag{8}$$

$$\langle \sigma_i \sigma_j \rangle = \sum_{\sigma} \sigma_i \sigma_j \mathbf{P}(\sigma) \tag{9}$$

As observed, it's required to find suitable parameters, $\vec{x} = \{h_i, J_{ij}\}$ that minimize $L(h_i, J_{ij})$. This is achievable by adopting a simple *gradient descent* algorithm. The algorithm deals with this optimization problem by taking steps in the parameter space, in the direction of $-\nabla f$, where f is the function to be minimized. In our case, the steps can be written as

$$\vec{x}_{n+1} = \vec{x}_n - \gamma \nabla L(\vec{x}_n) \tag{10}$$

where, for simplicity, the step size parameter γ is given a constant value, $\gamma = 0.1$. The gradient ∇L, is given by eqs. (6-7).

Provided with some start values $\vec{x}_0 = \{h_i^0, J_{ij}^0\}$, the algorithm begins to follow the negative gradient direction, bringing $\vec{x}$ closer to the local minimum through a stepping process, as schematically illustrated in figure 3.

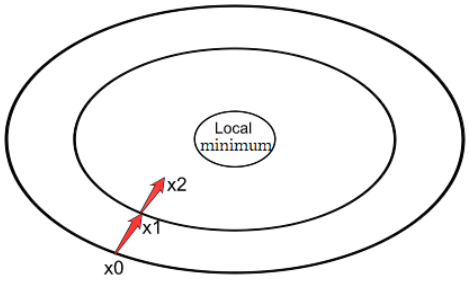

Figure 3: The gradient descent method. Starting from some point $\vec{x}_0$, and iteratively steps towards a local minimum of the negative log-likelihood function, L. The steps are perpendicular to the lines of constant L.

Let N_{par} denote the total number of parameters. The process is then iterated until the following stopping criterion is fulfilled

$$|\nabla L(\vec{x}_{n+1})| < \epsilon \sqrt{N_{par}} \tag{11}$$

with $\frac{|\nabla L(\vec{x}_{n+1})|}{\sqrt{N_{par}}}$ being the *root-mean-square deviation* (RMSD) of the model averages $\langle \sigma_i \rangle_{\vec{x}}$ and $\langle \sigma_i \sigma_j \rangle_{\vec{x}}$ from the input data $\langle \sigma_i \rangle^B$ and $\langle \sigma_i \sigma_j \rangle^B$. ϵ is a tolerance parameter, taking the value 0.01 in our calculations.

In the above procedure, the model averages $\langle \sigma_i \rangle_{\vec{x}_n}$ and $\langle \sigma_i \sigma_j \rangle_{\vec{x}_n}$ have to be computed for many different $\vec{x}_n$. This can in principle be done via the exact eqs. (8-9), although it can become computationally costly, unless the sequences are short. For N=30, the number of terms in the sums is $2^{30} \approx 10^9$. Considering this, a *Monte Carlo* (MC) simulation, based on the *Metropolis* algorithm, is used to estimate the averages.

Let $P_t(\boldsymbol{\sigma})$ denote the probability of observing $\boldsymbol{\sigma}$ after t steps. The MC algorithm generates a Markov chain such that

$$P_{t+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma'}} W(\boldsymbol{\sigma'} \to \boldsymbol{\sigma}) P_t(\boldsymbol{\sigma'}) \tag{12}$$

where the $W(\boldsymbol{\sigma'} \to \boldsymbol{\sigma})$'s are the transition probabilities (conditional probabilities of having the system in $\boldsymbol{\sigma'}$ at $t+1$ given that it is in $\boldsymbol{\sigma}$ at t). The transition probabilities are chosen such that (independently of what $P_{t=0}(\boldsymbol{\sigma})$ is)

$$\lim_{t \to \infty} P_t(\boldsymbol{\sigma}) = P(\boldsymbol{\sigma})$$

where $P(\boldsymbol{\sigma})$ denotes the desired probability distribution in eq. (1).

This property is achieved if the following conditions are satisfied:

1. Each state is accessible from any other state (ergodicity).
2. *Detailed balance* is fulfilled.

Here condition 2 implies that

$$W(\boldsymbol{\sigma} \to \boldsymbol{\sigma'})P(\boldsymbol{\sigma}) = W(\boldsymbol{\sigma'} \to \boldsymbol{\sigma})P(\boldsymbol{\sigma'})$$

The *Metropolis* algorithm proposes a new state of the system, which is either accepted or rejected. Assuming that the probabilities of proposing $\boldsymbol{\sigma} \to \boldsymbol{\sigma'}$ and $\boldsymbol{\sigma'} \to \boldsymbol{\sigma}$ are the same (symmetric), condition 2 can be satisfied if the acceptance probability, P_{acc}, is chosen as

$$P_{acc}(\boldsymbol{\sigma} \to \boldsymbol{\sigma'}) = \min\left\{1, \frac{P(\boldsymbol{\sigma'})}{P(\boldsymbol{\sigma})}\right\} \tag{13}$$

We implement this algorithm through the following steps:

1. Select a start configuration.
2. Perform a random trial change in the configuration.
3. Compute $\frac{P(\boldsymbol{\sigma'})}{P(\boldsymbol{\sigma})}$ where $\boldsymbol{\sigma'}$ and $\boldsymbol{\sigma}$ are the new and old configurations, respectively.
4. If $1 \leq \frac{P(\boldsymbol{\sigma'})}{P(\boldsymbol{\sigma})}$, accept the new configuration and go to step 6.
5. Else, pick a uniform random number R from the interval [0,1]. If $R \leq \frac{P(\boldsymbol{\sigma'})}{P(\boldsymbol{\sigma})}$, accept the new configuration. If the proposed change is not accepted, the old configuration is kept.
6. Repeat the steps 2-5 to collect a decent number of configurations.
7. Estimate the averages and related errors of the configurations

The implemented steps 4 and 5 are based on eq. (13).

The methods described above are applied to three HP structures with $N = 16$, $N = 25$ and $N = 30$, respectively. These structures are extracted from a database [12], which contains all sequences of length $N \leq 30$ that possess a unique minimum-energy structure, along with their structures. In the database, the structures are given in a four-letter alphabet: U (step up), D (down), R (right) and L (left). These letters act as a blue print for drawing the structure. The number of sequences associated with a given structure is denoted by B. To test DCA, the structure with the highest B is chosen for $N = 16$ (HPN16), $N = 25$ (HPN25) and $N = 30$ (HPN30). These structures have $B = 26$, $B = 326$ and $B = 813$, respectively. Averages in the statistical model are estimated using the Metropolis algorithm, implemented in a C-program. The empirical data required as input are calculated using *Python*.

5 Preparation of Input Data: A Simple Example

As input, the above procedure needs the quantities $\langle \sigma_i \rangle^B$ and $\langle \sigma_i \sigma_j \rangle^B$, which are averages over database sequences. To illustrate the calculation of these averages, consider the following $N = 4$ example:

$$\boldsymbol{\sigma} = \begin{cases} HPPH \\ HPHH \\ HHPH \\ HHHH \end{cases}$$

Since these four sequences ($B = 4$) are known to share the fold shown and already are aligned, we can calculate the empirical $\langle \sigma_i \rangle^B$ and $\langle \sigma_i \sigma_j \rangle^B$ through the scoring scheme illustrated in figure 4.

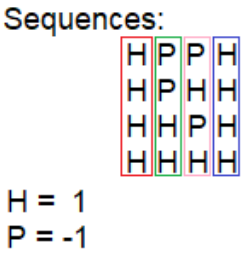

Figure 4: Different colours indicate different sequence positions. A hydrophobic residue (H) has a scoring point 1, and a polar residue (P) has -1.

The quantities $\langle \sigma_i \rangle^B$, given by eq. (2), are averages for the different columns in figure 1 (red, green, pink, and purple). Taking the scoring points into account, gives the red column a 1 , green 0, pink 0, and purple 1 as demonstrated in Table 1.

Table 1: Calculation of $\langle \sigma_i \rangle^B$.

Column	Calculation	Outcome
Red	$\frac{1}{4}(1+1+1+1)$	1
Green	$\frac{1}{4}(-1-1+1+1)$	0
Pink	$\frac{1}{4}(-1+1-1+1)$	0
Purple	$\frac{1}{4}(1+1+1+1)$	1

To compute $\langle \sigma_i \sigma_j \rangle^B$, each amino acid's scoring point in a column, will be multiplied by the scoring point of other amino acids sitting in the same row but in a different column, to be later added together according to (3). Calculations are illustrated in Table 2.

Table 2: Calculation of $\langle \sigma_i \sigma_j \rangle^B$.			
Column	Column	Calculation	Outcome
Red	Green	$\frac{1}{4}(1\cdot(-1)+1\cdot(-1)+1\cdot(1)+1\cdot(1))$	0
Red	Pink	$\frac{1}{4}(1\cdot(-1)+1\cdot(1)+1\cdot(-1)+1\cdot(1))$	0
Red	Purple	$\frac{1}{4}(1\cdot(1)+1\cdot(1)+1\cdot(1)+1\cdot(1))$	1
Green	Pink	$\frac{1}{4}(-1\cdot(-1)-1\cdot(1)+1\cdot(-1)+1\cdot(1))$	0
Green	Purple	$\frac{1}{4}(-1\cdot(1)-1\cdot(1)+1\cdot(1)+1\cdot(1))$	0
Pink	Purple	$\frac{1}{4}(-1\cdot(1)+1\cdot(1)-1\cdot(1)+1\cdot(1))$	0

The total number of $\langle \sigma_i \sigma_j \rangle^B$ values is given by $\frac{N(N-1)}{2}$, where N stand for the number of amino acids in the sequence, thus, in the case of figure 4, there are 6 $\langle \sigma_i \sigma_j \rangle^B$ values (N= 4). However, as a consequence of parity, only the result of the highlighted row is considered.

For more and longer protein sequences, a manual estimation of $\langle \sigma_i \rangle^B$ and $\langle \sigma_i \sigma_j \rangle^B$ becomes intractable. The calculations presented below were performed using *Python*. The results were saved as a text file, to be used as input for the statistical model computations. For the model, we then try to find $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$ that reproduce the input data. This is achieved through applying a *Metropolis* algorithm and finding the $\{h_i, J_{ij}\}$ that minimizes the log-likelihood function, via a gradient descent methodology. When this is accomplished, the validity of the results given by the statistical model is examined, to check if DCA indicates any direct couplings between the amino acids.

6 Results

In the gradient descent search, the stopping criterion ϵ (eq. 11) was set to 10^{-2}. The number of steps taken by the gradient descent, n, until the stopping criterion is fulfilled for the $N = 16$, 25 and 30 test proteins are $n = 94$, 91 and 54 respectively.

6.1 N= 16 example

The $N = 16$ structure studied can be found in figure 5. There are $B = 26$ sequences with this structure as their unique minimum-energy state. For an $N = 16$ HP structure, there are 16 h_i parameters and 49 J_{ij} couplings to be determined. We want to find out whether or not a large absolute value of a derived coupling, $|J_{ij}|$, can be taken to indicate a small distance R_{ij} between the residues i and j. To address this question, figure 6 shows an R_{ij}, J_{ij} scatter plot, where each data point represents one residue pair i, j. The data suggests that a weak negative correlation between these two quantities exist, such that nearest-neighbors tend to have an elevated J_{ij}. In particular, the three highest J_{ij} values all correspond to nearest-neighbor pairs, which are indicated in figure 5. At the same time, there are large negative couplings that do not correspond to nearest-neighbor pairs. As a result, the possibilities to identify nearest-neighbor pairs based on $|J_{ij}|$ are limited.

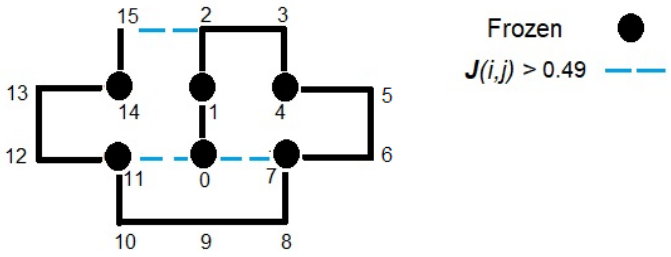

Figure 5: The $N = 16$ structure studied. Filled circles indicate conserved residue positions. The blue dashed lines indicate the three residue pairs with highest J_{ij} (> 0.49). These couplings are $J(0, 7) = 0.51$, $J(0, 11) = 0.50$ and $J(2, 15) = 0.74$.

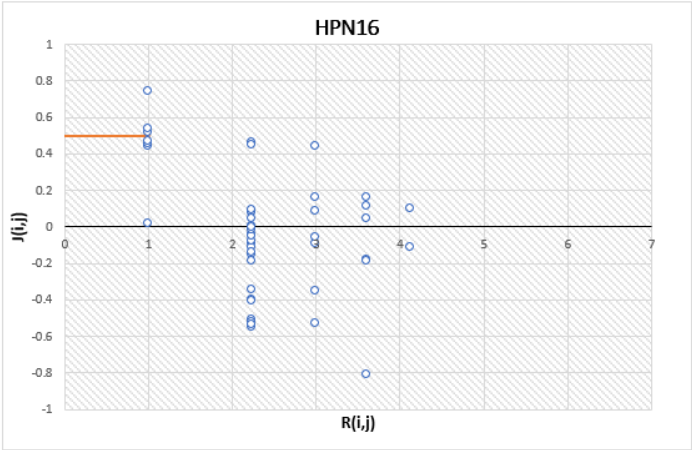

Figure 6: An R_{ij}, J_{ij} scatter plot, where R_{ij} is the distance and J_{ij} is the coupling between residues i and j, for $N = 16$. Each plot symbol represents one residue pair. The horizontal orange line is at $J_{ij} = 0.49$.

6.2 $N = 25$ example

The $N = 25$ structure studied can be found in figure 7. There are $B = 326$ sequences with this structure as their unique minimum-energy state. For an $N = 25$ HP structure, there are 25 h_i parameters and 132 J_{ij} couplings to be determined. In figure 8, an R_{ij}, J_{ij} scatter plot is shown, where each data point represents one residue pair i, j. The highest J_{ij} value corresponds to the nearest-neighbor pair, which is indicated in figure 7. At the same time, there are large negative couplings that do not correspond to nearest-neighbor pairs.

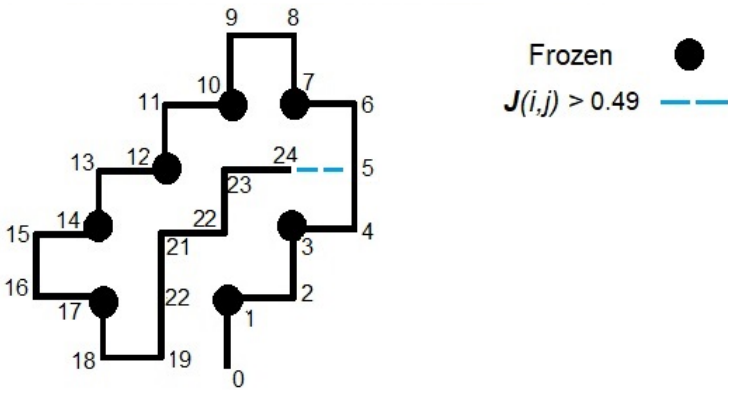

Figure 7: The $N = 25$ structure studied. Filled circles indicate conserved residue positions. The blue dashed lines indicate the residue pair with highest J_{ij} (> 0.49), that is $J(5, 24) = 0.75$.

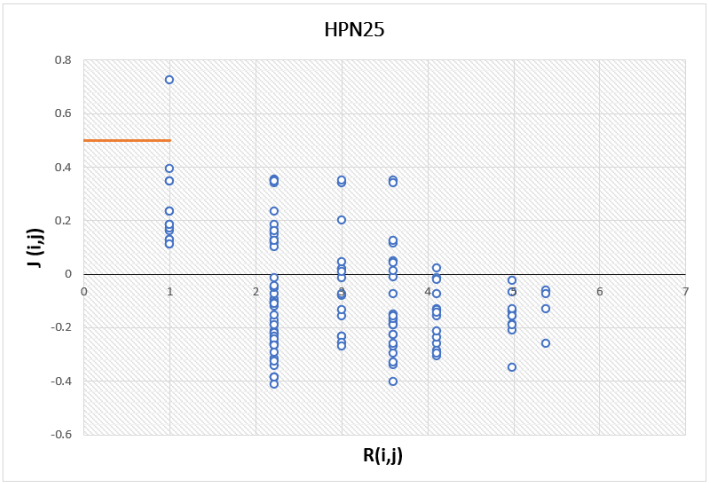

Figure 8: An R_{ij}, J_{ij} scatter plot, where R_{ij} is the distance and J_{ij} is the coupling between residues i and j, for $N = 25$. Each plot symbol represents one residue pair. The horizontal orange line is at $J_{ij} = 0.49$.

6.3 $N = 30$ example

The $N = 30$ structure studied can be found in figure 9. There are $B = 813$ sequences with this structure as their unique minimum-energy state. For an $N = 30$ HP structure, there are 30 h_i parameters and 196 J_{ij} couplings to be determined. In figure 10, an R_{ij}, J_{ij} scatter plot is shown, where each data point represents one residue pair i, j. The three highest J_{ij} values all correspond to the nearest-neighbor pairs, which are indicated in figure 9.

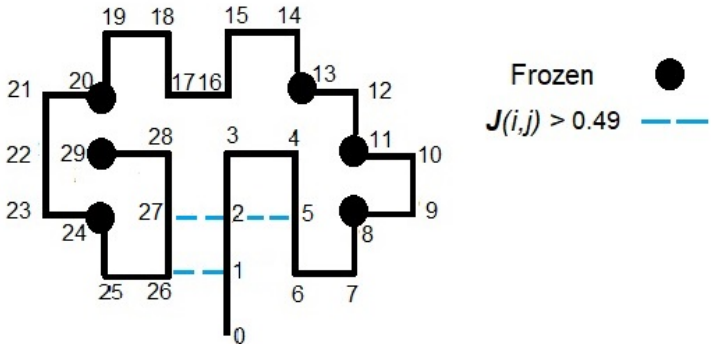

Figure 9: The $N = 30$ structure studied. Filled circles indicate conserved residue positions. The blue dashed lines indicate the three residue pairs with highest J_{ij} (> 0.49). These couplings are $J(1, 26) = 0.55$, $J(2, 5) = 0.53$ and $J(2, 27) = 0.74$.

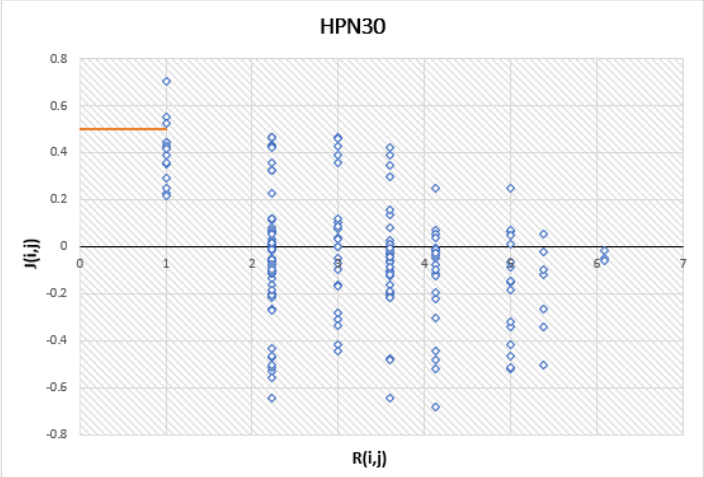

Figure 10: An R_{ij}, J_{ij} scatter plot, where R_{ij} is the distance and J_{ij} is the coupling between residues i and j, for $N = 30$. Each plot symbol represents one residue pair. The horizontal orange line is at $J_{ij} = 0.49$.

6.4 Uncertainties

In the procedure followed to minimize the negative log-likelihood function, L, there are two main potential sources of error. One is that the gradient descent search may identify a local minimum rather than the global one. The other is the statistical errors in the MC estimates. As for the gradient descent, different start values (seeds) have been considered, to check if new results could be obtained. No major change was observed in our results. The impact of statistical errors on the MC results was estimated. A *root-mean-square error*,

RMSE, was computed as

$$RMSE = \sqrt{\frac{1}{N_{par}} \sum_{i=1}^{N_{par}} \delta_i^2} \tag{14}$$

in which δ_i denotes the error in observable i, and N_{par} is the number of observables. RMSE was estimated to be roughly 10^{-4}.

To monitor convergence, the root-mean-square deviation, RMSD, statistical averages (given by eqs. (8-9)) and empirical averages (given by eqs. (2-3)) was evaluated. Figure 11 shows that RMSD decreases steadily with increasing n. It can't keep on decreasing forever, due

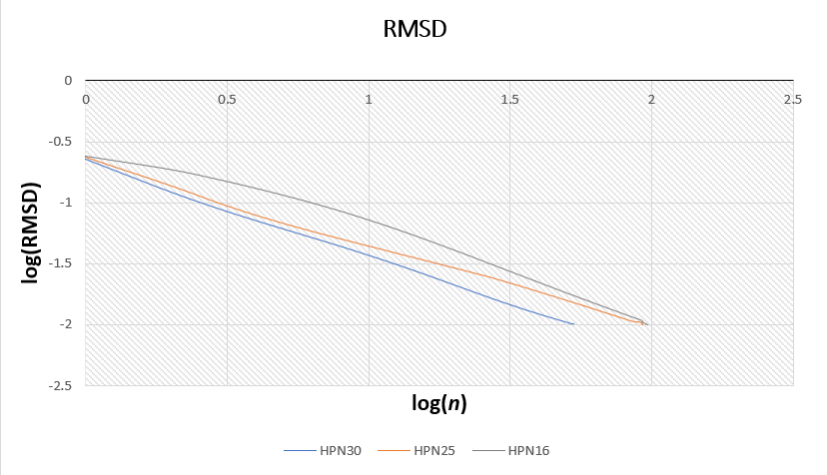

Figure 11: The RMSD between the statistical and empirical data converges to the chosen tolerance parameter, $\epsilon = 10^{-2}$, after a certain number of steps, n. Results for all our three examples are shown.

to the statistical MC errors. However, RMSE remains negligible when the RMSD reaches our stopping criterion (eq. (11)).

7 Summary and Discussion

The direct coupling analysis (DCA) method adopts statistical approaches based on inverse statistical mechanics, to predict contacts occurring between residues in a protein structure. The method is useful for real proteins, where the revealed contacts can facilitate structure prediction.

In this thesis, we have tested DCA on a simplified protein model, the HP model, in which proteins of length $N = 16$, 25 and 30 have been the test objects.

Solving the inverse statistical problem (ISP), for our simple model is relatively straightforward compared with solving it for real proteins. To deal with real proteins, systematic approximations are required. In our HP model, we take advantage of Monte Carlo calculations whith small statistical errors and no systematic errors to solve our ISP.

Regarding the produced results by DCA, we conclude, that having high and positive J_{ij} values, is a non trivial indication of having a direct coupling ($R_{ij} = 1$). At the same time, there exist some $J_{ij} \leq 0$ (shown in figures 6,8 and 10), such that their absolute value, $|J_{ij}|$, is higher than the largest positive J_{ij}, for i and j not being closest neighbours in the structure.

In conclusion, this may indicate that DCA is working less well, when applied to HP proteins than real proteins. Given that they represent a simplified model version of real proteins, this possibility can not be excluded. However, before drawing this conclusion, some potential issues should be investigated. One possible issue is failing to achieve the correct global minimum by the gradient descent method, since we have no guarantee of reaching the correct global minimum. Nevertheless, we have tried various random initial values (seeds), and repeated the stepping process. No major differences were noticed. Furthermore, the statistical and empirical averages agree well. Thus, we do not believe this is the issue.

Lastly, an additional matter we wanted to investigate further, but unfortunately couldn't due to limited time, is implementing a different scoring system, in which we use (0,1) instead of (-1,1) for polar (P) and hydrophobic (H) residues. We are curious as to whether this would improve the quality of our achieved results (e.g. finding more direct couplings).

References

[1] Anfinsen, C. B., 1973. Principles that govern the folding of protein chains. *Science* 181:7.

[2] Fang, Y., 2012. Protein Folding: The Gibbs Free Energy. *arXiv preprint arXiv:1202.1358* .

[3] Alberts, B., 2017. Molecular biology of the cell. Garland science.

[4] Finkelstein, A. V., and O. Ptitsyn, 2016. Protein physics: a course of lectures. Elsevier.

[5] Dill, K. A., 1990. Dominant forces in protein folding. *Biochemistry* 29:7133–7155.

[6] Kauzmann, W., 1959. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry* 14:1–63.

[7] Dill, K. A., 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24:1501–1509.

[8] Wüthrich, K., 2001. The way to NMR structures of proteins. *Nature Structural & Molecular Biology* 8:923–925.

[9] Ekeberg, M., C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* 87:012707–3.

[10] Morcos, F., T. Hwa, J. N. Onuchic, and M. Weigt, 2014. Direct coupling analysis for protein contact prediction. Springer.

[11] Lau, K. F., and K. A. Dill, 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997.

[12] Holzgräfe, C., A. Irbäck, and C. Troein, 2011. Mutation-induced fold switching among lattice proteins. *The Journal of Chemical Physics* 135:195101.

[13] Bahi, J. M., C. Guyeux, J.-M. Nicod, and L. Philippe, 2013. Protein structure prediction software generate two different sets of conformations. Or the study of unfolded self-avoiding walks. *arXiv preprint arXiv:1306.1439* .

[14] Böckenhauer, H.-J., and D. Bongartz, 2007. Protein folding in the HP model on grid lattices with diagonals. *Discrete Applied Mathematics* 155:230–256.

[15] Cocco, S., R. Monasson, and M. Weigt, 2013. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS computational biology* 9:e1003176.

[16] Ekeberg, M., C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* 87:012707–4.