



**LUNDS**  
UNIVERSITET

DERIVATION OF A REWEIGHTING ALGORITHM  
FOR THE PREDICTION OF THE  
THERMODYNAMIC STABILITY OF MUTATED  
SEQUENCES AND EVALUATION OF ITS  
USABILITY IN A SIMPLE CONTINUOUS PROTEIN  
MODEL

Frieder Henning

Supervisor: Stefan Wallin

A thesis submitted for the degree of  
Bachelor of Science

January 2015

## Abstract

In this thesis, a new reweighting algorithm for predicting the native conformations of amino acid sequences, consisting of 16 residues in a reduced representation continuous protein model, is derived and tested to map neutral nets of  $\alpha$ -helices and  $\beta$ -sheets.

For examining the potential of the reweighting algorithm for predicting the thermodynamic stability of amino acid sequences, it is applied to a mutational pathway, connecting a  $\alpha$ -helix to a  $\beta$ -sheet. Moreover, a bistable sequence is used to predict the evolution of the chain properties along the pathway. The reweighted averages of four relevant observables are then compared to the averages obtained in direct simulations. Further, neutral nets of  $\alpha$ -helices and  $\beta$ -sheets are mapped by using the reweighting algorithm. The compositions of the neutral nets are analyzed and the validity of the mapping is tested, both, in direct simulations and the analysis of a few sequences.

The evaluation of the usability of the reweighting algorithm indicates, that, in principle, it is capable of predicting the stability of native conformations. Moreover, the test cases show that for this method, bistable sequences are the most appropriate starting point for making predictions of the stability of other sequences. In the analysis of the neutral nets, it is shown that the compositions of the nets display some realistic features. However, as it turns out, some sequences, that are identified as belonging to either of the two nets do not fulfill the requirements that define these nets.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Proteins and Amino Acids . . . . .	3
2.2	Protein Folding . . . . .	4
2.3	Protein Evolution and Mutations . . . . .	6
<b>3</b>	<b>Model and Methods</b>	<b>7</b>
3.1	The Model . . . . .	7
3.2	Energy Contributions . . . . .	8
3.3	Metropolis Algorithm . . . . .	9
3.4	Reweighting Method . . . . .	10
3.5	Application . . . . .	10
3.6	Observables . . . . .	12
3.7	Simulations . . . . .	13
<b>4</b>	<b>Results</b>	<b>14</b>
4.1	Mutational Pathway . . . . .	14
4.2	Validity of Reweighting for Different Observables . . . . .	15
4.3	Bistable Sequence . . . . .	19
4.4	Neutral Net . . . . .	21
<b>5</b>	<b>Discussion</b>	<b>25</b>
<b>6</b>	<b>Conclusions</b>	<b>28</b>

# 1 Introduction

Proteins are biological macro-molecules that are made up of amino acids. They can be found in all cells and carry out many exercises that are essential for life as we know it. As enzymes, proteins catalyze reactions in cells and are involved in e.g. metabolism or the creation of deoxyribonucleic acid (DNA). Molecular motors that transport molecules within cells are built up of proteins and as signaling molecules, proteins are involved in inter-cellular communication [1].

For a proper understanding of how proteins can execute these different tasks it is insufficient to know the constituting amino acid sequences. This is because the function of proteins is dependent on their three-dimensional structure, which is the result of a process called protein folding.

For the study of protein folding and understanding of the mechanisms that lie behind this process, various computer models have been developed. These allow to make predictions of the three-dimensional structure of proteins, based on their amino acid sequences. Simulations of these models can be computationally demanding. In order to be able to execute extensive investigations of protein properties, simple models that only capture essential protein properties have been developed. However, even with simple models, theoretical studies are limited by the extremely large sequence space.

In this project, an algorithm that can predict the thermodynamic behavior such as the stability of the three-dimensional structure of an amino acid sequence,  $\hat{\sigma}$ , given the simulation data from another sequence  $\sigma$ , is derived and tested. The algorithm is based on so-called histogram reweighting techniques [13]. After the derivation and testing, the algorithm is applied to a biophysical problem.

In the second section of this thesis, elementary background information on proteins, protein folding and protein evolution is given. It helps to understand the assumptions which the model that is used in this project is based on and it is needed to understand the significance of the applications in the latter part of this thesis.

The third section presents the applied model in greater detail. The fundamental ideas and necessary simplifications are explained. Furthermore, the theory for the implemented reweighting algorithm is derived and the observables, that are relevant in this project are introduced. In the last part of this section, it is briefly described how the data are sampled.

The fourth section is dedicated to the results that are obtained after the application of the reweighting method. The evolution of an amino acid sequence is analyzed. Predictions about the properties of the proteins are tested. Moreover, the sequence space of two structures is studied and composition of sequences that arrange into these structures is investigated.

Finally the results are discussed concerning their reliability and proposals for further improvements, regarding the methodology, are given.

## 2 Background

In this section, the different levels on which proteins can be described are introduced. Starting from linear amino acid sequences, the concept of protein folding and the three-dimensional structures that this process results into are introduced. In order to understand the folding dynamics, the most important interactions, that give rise to the process are described. Finally, some background on protein evolution is given.

### 2.1 Proteins and Amino Acids

Proteins are polypeptides, i.e. chains that are made up of smaller subunits, the so called peptides. The building blocks of these peptides in turn are amino acids. There are twenty-three different amino acids occurring in nature, but usually only twenty of them are incorporated in mammalian proteins [16].

The basic architecture is the same for all amino acids. In the middle of every amino acid sits the backbone. This is a carbon atom (the so called  $\alpha$ -carbon) connected to an hydrogen atom and some side group. Moreover, the  $\alpha$ -carbon is joined to a carboxyl group (-COOH) and to an amino group (-NH<sub>2</sub>). The carbon atom of the carboxyl group is sometimes denoted as C'. A general plan of the architecture can be seen in figure 1 [16].

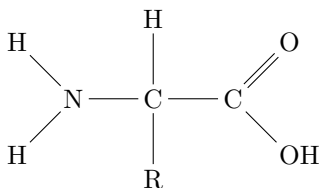


Figure 1: The basic architecture of an amino acid. In the center sits the  $\alpha$ -carbon. The amino group can be seen to the left and the carboxyl group to the right. The side group that is connected to the  $\alpha$ -carbon and determines the identity of the amino acid is denoted by R.

The side group determines the identity of any amino acid as well as its physical properties. Roughly speaking, amino acids can be divided into three classes. Dependent on the side chain, they can be either polar or hydrophobic. A third special case includes two amino acids only, namely glycine and proline. These amino acids are special due to their geometrical properties. They are often found in turns of the three-dimensional structure of proteins [8].

In protein synthesis, amino acids are connected to each other in the condensation reaction. In this reaction the carboxyl group of one amino acid is connected to the amino group of another amino acid. As a by-product, a H<sub>2</sub>O-molecule is released. The covalent bond between the two amino acids is called peptide bond. Exemplary figure 2 shows a sketch of two amino acids bonded to each

other by a peptide bond [16].

Typically, a simple protein consists of 30-400 building blocks. Its sequence of amino acids makes up the so called primary structure [16].

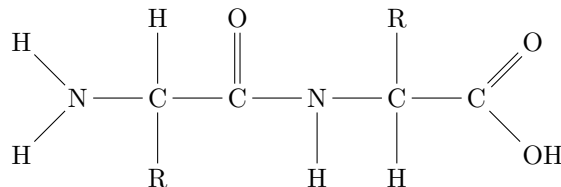


Figure 2: Two amino acids connected to each other by a peptide bond. In this reaction, the HO that is released from the carboxyl group and the H atom, that is released from the nitrogen, bind to each other and leave as a H<sub>2</sub>O-molecule.

## 2.2 Protein Folding

For a proper description of a protein, its physical shape is of extraordinary importance as it determines the protein's functionality [8]. The equilibrium shape of a protein is a direct consequence of the mutual interactions between the constituting residues and hence a consequence of the primary structure. In equilibrium, the protein is folded into the state of minimum free energy, the so called native conformation. This native conformation is uniquely determined by the amino acid sequence. However, its stability is dependent on the proteins environment, i.e. the temperature, pH or solvent have influence on the equilibrium shape of the protein [9].

When only some short sequence of residues and the structure they fold into is considered, the resulting three-dimensional arrangement is called secondary structure. Hence, the secondary structure describes the local folding behavior of an amino acid sequence. For simple proteins, the sum of all secondary structures makes up the tertiary structure, i.e. the shape all residues arrange into in the respective environment [16].

Two common types of secondary structures are the  $\alpha$ -helix and the  $\beta$ -sheet. As the name implies, in an  $\alpha$ -helix the amino acids are arranged in an helical structure. Typically, this structure is right-handed. Moreover, each amino acid makes a turn of  $100^\circ$  in this arrangement and hence there are 3.6 amino acids involved in one revolution around the helical axis. Another characteristic of the  $\alpha$ -helix are the hydrogen bonds that connect the amide group of one amino acid with the carbonyl group that is placed four positions earlier in the sequence. These bonds make the  $\alpha$ -helix a very stable structure [12].

Another regular structure, that is encountered frequently in proteins, is the so called  $\beta$ -sheet. In this structure the amino acids are arranged in parallel layers. Again, these layers are connected to each other by hydrogen bonds between the amide and carbonyl groups of amino acids that are situated opposite to each

other. When only two amino acid strands are considered this structure is called a  $\beta$ -hairpin [12].

Yet not every primary structure necessarily folds into some well defined secondary structure in any arbitrary environment. Some sequences will be found in random conformations most of the time. Another class of sequences are bistable. In the same environment, these sequences can fold into both,  $\alpha$ -helices and  $\beta$ -sheets [8].

In general, the relation between the primary and secondary structure is very complex, due to the different interactions that are involved in the folding process. To some extent, it is still unclear which interactions are involved in folding and how large the contributions of the different interactions to the structure formation are [9].

However, from experimental evidence it can be deduced that there are four interactions that are most important in the folding process. These are the hydrophobic effect, H-bonding between chain residues, electrostatic interactions between chain residues and van der Waals interactions. [9].

In an aqueous environment, the hydrophobic effect seems to be the major driving force behind protein folding. This effect is entropic in nature and stems from the interactions of the protein side chains with the water molecules of the solvent. As nonpolar side chains cannot build H-bonds with water molecules, they disturb the H-bond network of water in a way that decreases the entropy of the solvent/protein system and consequently even its free energy. To minimize this free energy contribution it is therefore energetically favorable to keep hydrophobic side chains and water molecules apart [9].

On the other hand, polar side chains do not disturb the H-bond network in the same way. In real proteins, the hydrophobic side chains are therefore often found in the core of the protein, whereas the polar side chains are exposed to the surrounding. Effectively this means that the polar side chains shield the hydrophobic side chains from interacting with the water molecules [9].

Although not a major driving force in protein folding, hydrogen bonds are involved in the formation of the secondary structure as well. In proteins, the most important hydrogen bonds are found between the carbonyl (C=O) and amide groups (NH) of the backbone. These bonds are especially important in the stabilization of the three-dimensional structure [9].

Moreover, electrostatic interactions between ions that are incorporated in the protein, as well as van der Waals interactions affect the stability of proteins. These interactions, however, are of minor importance for protein folding and can be repulsive or attractive, dependent on the ion or the pH of the solvent [9].

### 2.3 Protein Evolution and Mutations

The study of protein evolution is interesting for different reasons. On one hand, it does help to improve the understanding of how life could evolve. On the other hand, a better understanding of protein evolution and mutations is essential for e.g. potential medical applications [5].

Protein evolution is the consequence of mutations. There are different mutational mechanisms. In point mutations, one amino acid is replaced by another amino acid. Insertions and deletions are mutations where one residue of the polypeptide is simply inserted or removed from the chain respectively. Other forms of mutations include e.g. duplication or recombinations [7].

In protein evolution, the relation between primary and secondary structure is of crucial importance. Again, this is due to the relation between the shape of the protein and its functionality. It follows that mutations that change the three-dimensional structure of a protein can have potentially lethal consequences. However, it has been observed that often there is no change in the secondary structure upon a simple point mutation. In principle, this means that the functionality of the mutated protein is not affected. Mutations that do not change protein functionality are called neutral mutations. They give rise to neutral evolution. This neutral evolution, that is going on under the surface, can make sequences accessible that would not be accessible in a single point mutation from the original sequence otherwise [7].

In protein evolution modeling, the Hamming distance,  $H$ , has been introduced as a measure for the relation between two primary structures. It is defined as the number of positions in sequence space in which two sequences differ [5].

All sequences that have the same secondary structure and that are connected to each other by consecutive single point mutations are said to be part of one neutral net. The sequence that can undergo the most single point mutations within this neutral net is the so called prototype sequence [6].



## 3 Model and Methods

In this section, the fundamental assumptions, that the model is based on, are presented. This includes the geometry of the model chain, the mathematical description of the energy contributions as well as the Monte Carlo algorithm that is used to simulate the thermal motion of the model chain. The fundamental idea of reweighting methods is presented and applied to amino acid sequences. Moreover, the relevant observables that are studied in the further course of this project are introduced. Finally, the simulations and data sampling for this project are described.

### 3.1 The Model

The model that is used in this project is developed in Ref. [4]. This subsection as well as subsections 3.2 and 3.3 follow the discussion from this reference. The implementation of the model as a computer program has been performed in advance of the project described in this thesis. The model is a so called "reduced-representation, continuous protein model". The representation is reduced in the sense that there are three different amino acids considered only. Two of these amino acids are characterized by their hydrophobicity as they can either be hydrophobic (h) or polar (p). In the third case the geometrical properties and interaction behavior of the amino acid resemble those of glycine (t). In every amino acid the backbone is represented in atomic detail, whereas the side chain is reduced to an enlarged carbon atom. This enlargement is meant to mimic the bulkiness of the side-chain.

That the model is continuous, contrasts it from the simpler HP model. In the HP model, which is used to determine the minimum free energy conformation of some primary structure, the positions of the residues are restricted to be placed on the sides of some lattice. In the continuous model, the residues that are attached to each other are free to rotate.

Out of the three dihedral backbone angles,  $\omega$ , the angle, that defines the bond between the C' and N of the peptide bond, is fixed to be  $180^\circ$ . The other two angles  $\phi$  (the bond between N and  $C_\alpha$ ) and  $\psi$  (the bond between  $C_\alpha$  and C') are free. This means that there are  $2N$  degrees of freedom for a chain consisting of  $N$  amino acids. A set of  $2N$  angles defines the conformation of an amino acid chain. Equivalently, a conformation can be described in terms of the Cartesian coordinates of the atoms.

In the search for the secondary structure of an amino acid sequence, the thermal motion of the polypeptide is simulated. This is done by means of a Monte Carlo algorithm, the so called Metropolis algorithm. Whether transitions between different states are accepted depends on some random component, but first and foremost it depends on the energy difference between the states. Four different energy terms contribute to the conformational energy.

## 3.2 Energy Contributions

In the model, protein folding is linked to four energy contributions. The energy of conformation  $C$  is given by the sum  $E(C) = E_{\text{exvol}} + E_{\text{local}} + E_{\text{hbond}} + E_{\text{hp}}$ .

The first term  $E_{\text{exvol}}$  is the excluded-volume energy. Mathematically, it is of the form,

$$E_{\text{exvol}} = k_{\text{exvol}} \sum_{i < j} \left( \frac{\lambda_{ij} \sigma_{ij}}{r_{ij}} \right)^{12}. \quad (1)$$

This is a sum over all atom pairs  $ij$ .  $\sigma_{ij}$  is the sum of the atomic radii  $\sigma_i + \sigma_j$  and  $r_{ij}$  is the distance between these atoms.  $\lambda_{ij}$  is a scale factor and  $k_{\text{exvol}}$  a weight factor respectively. The overall effect of this term is to make it energetically unfavorable for the atoms to occupy the same spatial coordinates.

The second energy term,  $E_{\text{local}}$ , represents the interaction of electric charges on the protein backbone,

$$E_{\text{local}} = k_{\text{local}} \sum_I \sum_{i < j} \frac{q_i q_j}{r_{ij}}. \quad (2)$$

Here, the first sum is taken over all  $N$  amino acids in the chain and the second sum goes over the atoms in the respective amino acids.  $q_i$  and  $q_j$  are the partial charges of the atoms and  $r_{ij}$  is the distance between them.

The third energy contribution to the conformational energy,  $E_{\text{hbond}}$ , is due to hydrogen bonding between the carbonyl and amide groups. It is given by,

$$E_{\text{hbond}} = k_{\text{hbond}} \sum_{ij} \gamma_{ij} \left[ 5 \left( \frac{\sigma_{\text{hb}}}{r_{ij}} \right)^{12} - 6 \left( \frac{\sigma_{\text{hb}}}{r_{ij}} \right)^{10} \right] (\cos \alpha_{ij} \cos \beta_{ij})^{1/2}. \quad (3)$$

Here,  $\alpha_{ij}$  is the N-H-O angle and  $\beta_{ij}$  is the H-O-C' angle. This energy contribution is only included if both  $\alpha_{ij} > 90^\circ$  and  $\beta_{ij} > 90^\circ$ . Moreover, it is required that the  $ij$  pairs are separated by at least two amide groups to be included.  $\gamma_{ij}$  is some sequence dependent scale factor. The function of this term is to make hydrogen bonding with glycine energetically less favorable. This is to mimic the effect of glycine to break the secondary structure.

Finally, the fourth energy term,  $E_{\text{hp}}$ , stands for the hydrophobic effect. This energy contribution is given by,

$$E_{\text{hp}} = -k_{\text{hp}} \sum_{ij} e^{-\frac{(r_{ij} - \sigma_{\text{hp}})^2}{2}}. \quad (4)$$

Only amino acids with hydrophobic side chains that are separated by three or more side chains can contribute to this energy.

### 3.3 Metropolis Algorithm

The thermal motion of the protein is simulated using an extension of the Metropolis algorithm. The Metropolis algorithm is a Markov chain Monte Carlo method. It works under the assumption that the future development of a simulated system is independent of its history. Moreover, it is implemented to generate a distribution of states that follows the Boltzmann distribution [15].

In order to make the collection of states follow the Boltzmann distribution, the Markov chain, that is uniquely determined by its transition probability  $W(r \rightarrow r')$ , needs to fulfill two requirements [3]. These are,

1. The limit distribution is required to be stationary, i.e. if  $P^{(n)}$  is the limit distribution, it is required that  $P^{(n+1)} = P^{(n)}$
2. The limit distribution has to be unique, i.e. the transition probabilities need to be ergodic. Ergodicity means that the process is aperiodic and positive recurrent, i.e. the number of steps to go back to some state in conformation space has to be finite.

The first condition can be fulfilled by requiring detailed balance [17]. This means that in equilibrium, transitions from state  $r$  to state  $r'$  are compensated for by transitions in the reverse direction. Mathematically, this is expressed as,

$$P(r)W(r \rightarrow r') = P(r')W(r' \rightarrow r) , \quad (5)$$

which can be rewritten as,

$$\frac{W(r \rightarrow r')}{W(r' \rightarrow r)} = \frac{P(r')}{P(r)} . \quad (6)$$

The transition probability can be separated into two parts,

$$W(r \rightarrow r') = g(r \rightarrow r')A(r \rightarrow r') . \quad (7)$$

The function  $g(r \rightarrow r')$  is the so called proposal probability and is required to be ergodic. The function  $A(r \rightarrow r')$  is the acceptance probability and has to be chosen such that it fulfills the requirement of detailed balance. Substitution of (7) into (6) yields,

$$\frac{A(r \rightarrow r')}{A(r' \rightarrow r)} = \frac{P(r')}{P(r)} \frac{g(r' \rightarrow r)}{g(r \rightarrow r')} . \quad (8)$$

One acceptance probability that fulfills this requirement is the Metropolis choice,

$$A(r \rightarrow r') = \min \left( 1, \frac{P(r')}{P(r)} \frac{g(r' \rightarrow r)}{g(r \rightarrow r')} \right) . \quad (9)$$

For proposal probabilities that fulfil  $g(r \rightarrow r') = g(r' \rightarrow r)$ , and the choice  $P \propto \exp(-\beta E)$ , where  $E$  is the energy of the respective state and  $\beta = 1/k_B T$ , where  $k_B$  is the Boltzmann constant, equation (9) can be rewritten as,

$$A(r \rightarrow r') = \min (1, e^{-\beta \Delta E}) . \quad (10)$$

The motion of the residues is now simulated in two steps. In the first step a new conformation is generated. The dihedral angles, or equivalently the spatial coordinates, of this conformation are chosen randomly. Now, the energy difference,  $\Delta E$ , between the new conformation and the old conformation is calculated. For the case that  $\Delta E \leq 0$ , i.e. the new conformation is energetically favorable or equivalent, the new conformation is always accepted. If  $\Delta E > 0$ , a random number between 0 and 1 is generated and compared to  $\exp(-\beta\Delta E)$ . The new conformation is accepted if the random number is less than  $\exp(-\beta\Delta E)$  and rejected otherwise.

The extended Metropolis algorithm that is used in this model samples the behavior of a Markov chain at different temperatures, so-called simulated tempering. This does not only allow to study the behavior of a sequence at several different temperatures but reduces the correlation between the measurements taken at the same temperature as well. Further theory on this method can be found in Ref. [14].

### 3.4 Reweighting Method

Reweighting methods are a class of schemes that use probability distributions of states that are sampled in Monte Carlo simulations to predict the behavior of these Monte Carlo systems under conditions deviating from those that were assumed in the original simulation [13].

In this project a reweighting algorithm that is based on the theory presented in reference [2] is derived. Originally, this theory was developed to compute the free energy difference between two Monte Carlo systems but it can also be applied to analyze thermodynamic properties other than  $F$  [2].

In order to make predictions of the behavior of a system,  $S_0$ , in histogram reweighting, the sampling of the configuration space of another, comparable system  $S_1$  is used [10]. The thermodynamic properties of system  $S_1$  follow some distribution function. Knowledge of this distribution can be exploited to estimate the behavior of  $S_0$ . For instance, in the simulation of  $S_1$ , different configurations are visited and their corresponding energy can be calculated. The energy of the visited configurations can now be calculated for  $S_0$  and the energy difference  $\Delta E$  can be computed. This energy difference in turn is then used to construct a histogram over the probability density of the different  $\Delta E$  [10].

The main limitation of the method is that the overlap of the distribution in configuration space between the two systems needs to be large enough [2].

### 3.5 Application

In the following, a reweighting algorithm for predicting the effect of sequence changes on the thermodynamic properties of amino acid sequences is derived. However, the derivation is equally valid for the other observables that will be studied later in this thesis. The implementation of the developed theory has

been performed as a part of this project.

The probability of observing a particular state or conformation  $X$  with energy  $E(X)$  for sequence  $\sigma$  is given by the Boltzmann distribution,

$$p(X|\sigma) = \frac{1}{Z(\sigma)} e^{-\beta E(X, \sigma)} , \quad (11)$$

where  $Z(\sigma)$  is the partition function which is defined as,

$$Z = \sum_X e^{-\beta E(X)} . \quad (12)$$

In order to relate  $p(X|\sigma)$  to the probability of observing conformation  $X$  in another sequence,  $\hat{\sigma}$ , equation (11) is multiplied by  $e^{\beta E(X, \hat{\sigma})} \cdot e^{-\beta E(X, \hat{\sigma})} = 1$ . This yields,

$$p(X|\sigma) = \frac{1}{Z(\sigma)} e^{-\beta E(X, \hat{\sigma})} e^{\beta \Delta E} , \quad (13)$$

where  $\Delta E = E(X, \hat{\sigma}) - E(X, \sigma)$ . Another multiplication by  $Z(\hat{\sigma})/Z(\hat{\sigma}) = 1$ , allows for the identification of  $p(X|\hat{\sigma})$ , the probability of observing conformation  $X$  for sequence  $\hat{\sigma}$ . After multiplication one obtains,

$$p(X|\sigma) = \frac{Z(\hat{\sigma})}{Z(\sigma)} e^{\beta \Delta E} p(X|\hat{\sigma}) , \quad (14)$$

or equivalently,

$$p(X|\hat{\sigma}) = \frac{Z(\sigma)}{Z(\hat{\sigma})} e^{-\beta \Delta E} p(X|\sigma) . \quad (15)$$

The fraction  $Z(\hat{\sigma})/Z(\sigma)$ , in turn, can be rewritten and related to  $p(X|\hat{\sigma})$ . Using the definition of  $Z$ , (12), one gets,

$$\frac{Z(\hat{\sigma})}{Z(\sigma)} = \frac{1}{Z(\sigma)} \sum_X e^{-\beta E(X, \hat{\sigma})} = \frac{1}{Z(\sigma)} \sum_X e^{-\beta E(X, \sigma)} e^{-\beta \Delta E} . \quad (16)$$

In this expression the probability  $p(X|\hat{\sigma})$  can be identified and it can be written as,

$$\frac{Z(\hat{\sigma})}{Z(\sigma)} = \sum_X e^{-\beta \Delta E} p(X|\sigma) = \langle e^{-\beta \Delta E} \rangle_\sigma , \quad (17)$$

where  $\langle e^{-\beta \Delta E} \rangle_\sigma$  is the thermodynamic average obtained for sequence  $\sigma$ . Combining equations (15) and (17), one can see that the average of an observable,  $O$ , for sequence  $\hat{\sigma}$  can be obtained from a simulation of sequence  $\sigma$ . By definition the average of an observable is calculated according to,

$$\langle O \rangle_{\hat{\sigma}} = \sum_X O(X) p(X|\hat{\sigma}) , \quad (18)$$

but,

$$p(X|\hat{\sigma}) = \frac{1}{\langle e^{-\beta \Delta E} \rangle_\sigma} \sum_X O(X) e^{-\beta \Delta E} p(X|\sigma) , \quad (19)$$

and therefore,

$$\langle O \rangle_{\hat{\sigma}} = \frac{1}{\langle e^{-\beta \Delta E} \rangle_{\sigma}} \sum_X O(X) e^{-\beta \Delta E} p(X|\sigma) . \quad (20)$$

In terms of the parameter,

$$w = e^{-\beta \Delta E} , \quad (21)$$

equation (20) can be written as

$$\langle O \rangle_{\hat{\sigma}} = \frac{\langle Ow \rangle_{\sigma}}{\langle w \rangle_{\sigma}} , \quad (22)$$

The effect of the weighting parameter  $w$  is to increase the influence of those conformations with with large negative  $\Delta E$  on the total average  $\langle O \rangle_{\hat{\sigma}}$ .

### 3.6 Observables

In the model, a few observables are measured for every visited conformation. In this project not all of them are of interest and in the following, the focus lies on the behavior of only four observables. These are the energy,  $E$ , the end-to-end radius,  $R_{ee}$ , the  $\alpha$ -content and the  $\beta$ -content of the conformations.

The energy is considered as it is the starting point for the reweighting method and ultimately, it is the driving parameter behind protein folding. It has to be emphasized, that the parameter  $E$  is the sum of four energy contributions.

The end-to-end distance is defined by,

$$R_{ee} = |\mathbf{r}_1 - \mathbf{r}_N| . \quad (23)$$

where  $\mathbf{r}_1$  and  $\mathbf{r}_N$  are the  $\alpha$ -carbon positions of the first and last amino acids respectively.

It contains information about the geometrical shape of the protein. As the shape plays an important role for the functionality of the protein, this parameter is of interest.

The  $\alpha$ - and  $\beta$ -content measure the number of amino acids that are arranged in an  $\alpha$ -helix or  $\beta$ -sheet respectively. As the first and last links in the chain are more flexible, they rarely become part of these structures and are ignored when the  $\alpha$ - and  $\beta$ -content are measured. Hence, in simulations of chains consisting of  $N = 16$  amino acids the values of these observables can assume any integer value between 0 and 14.

Whether or not an amino acid is part of an  $\alpha$ -helix or  $\beta$ -sheet is determined by the dihedral backbone angles  $\phi_i$  and  $\psi_i$ .

### 3.7 Simulations

In this project, amino acid sequences of length  $N = 16$  are studied. The simulations of these sequences consist of  $10^7$  Monte Carlo cycles. In every cycle, 100 conformations are generated. The conformations were saved every thousandth Monte Carlo cycle. As in one run the behavior of the sequences is tested at eight different temperatures, only about one eighth of the saved conformations can be used for the calculation of the reweighted averages. Accordingly, these calculations are based on roughly 1250 conformations.

The typical duration of such simulations is one day. The duration for predictions using the histogram reweighting algorithm is typically less than twenty seconds.

For the estimate of uncertainties in the simulations and reweighted averages a Gaussian distribution of outcomes around the average values is assumed.

## 4 Results

The initial goal of this section is to test the accuracy and limitations of the implemented algorithm. For this purpose, the algorithm is applied to a mutational pathway that connects two sequences, A1 and N1, which fold into two different native states. The averages of the four observables obtained in these simulations are then compared to the predictions made by the reweighting algorithm. The algorithm is then tested for a bistable sequence that is unrelated to the mutational pathway. Finally, the neutral net of the two different structures are mapped out and an attempt of a first analysis of the neutral nets is made.

### 4.1 Mutational Pathway

The mutational pathway that is studied in this section is taken from reference [11] and shown in table 1. The respective end sequences A1 and N1 are designed such that they differ in their secondary structure. The native conformation of sequence A1 is an  $\alpha$ -helix whereas sequence N1 folds into a  $\beta$ -sheet. The minimum energy conformations for both sequences can be seen in figure 3.

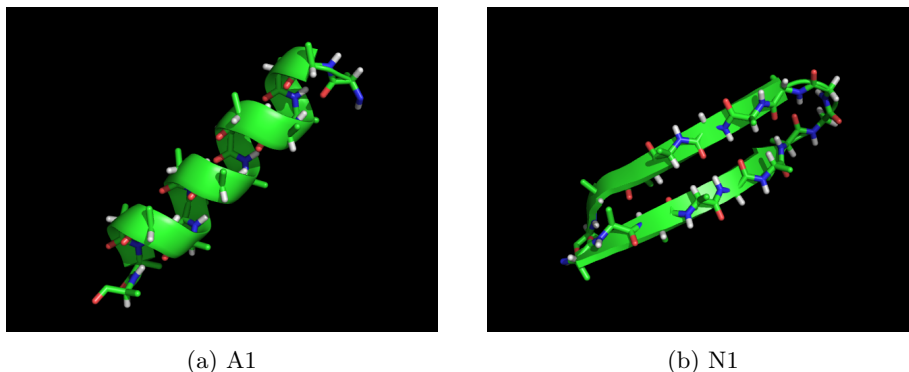


Figure 3: The minimum energy conformations of the sequences A1 and N1. Typically, the amino acids placed in the end of the sequences are not included in the  $\alpha$ -helix and  $\beta$ -sheet respectively.

The thermodynamic behavior of the sequences along the mutational pathway is measured in four independent simulations. The averages of the four observables for the respective sequences obtained in these simulations can be seen in figure 4.

The analysis of the averages shows clearly how the evolution from  $\alpha$ -helix to  $\beta$ -sheet takes place. The  $\alpha$ -content for sequences with Hamming distance  $H_{A1} \leq 7$  lies constantly at a high value of roughly 12 (figure 4a). The  $\beta$ -content of these sequences is very low and constantly at a value of around 0.3 (figure 4b). At Hamming distance  $H_{A1} = 8$  the  $\alpha$ -content decrease slightly whereas the  $\beta$ -content increases. A drastic change in the observables occurs between  $H_{A1} = 8$  and  $H_{A1} = 9$ . Here, the  $\alpha$ -content decreases from ten to one and the  $\beta$ -content increases from one to 6.5 respectively. This change in the observables represents the switch from an  $\alpha$ -helix to a  $\beta$ -sheet. Another slight change of the respective contents can again be measured between  $H_{A1} = 9$  and  $H_{A1} = 10$ . As the



Sequence	Name	$H_{A1}$	$H_{N1}$
phphphpttphphphp	N1	10	0
phphphptpphphphp	P1	9	1
phphphpppphphphp	P2	8	2
phpphpppphphphp	P3	7	3
phhpphpppphphphp	P4	6	4
phhpphppphhphphp	P5	5	5
phhpphpphhphppp	P6	4	6
phhpphphpphhphppp	P7	3	7
pphpphphpphhphppp	P8	2	8
pphpphphpphhphpp	P9	1	9
pphpphphpphphppp	A1	0	10

Table 1: The sequences that are studied in this project. The single point mutations that N1 undergoes in the evolution towards A1 are marked red. For convenience, the Hamming distances relative to A1 and N1, denoted by  $H_{A1}$  and  $H_{N1}$  respectively, are given in this table as well.

$\alpha$ -content of A1 is higher than the  $\beta$ -content of N1 it can be concluded that the  $\alpha$ -helix formed by A1 is more stable than the  $\beta$ -sheet that sequence N1 arranges into.

This switch is also directly reflected in the end-to-end distances that are measured along the mutational pathway (figure 4c). The native conformations of the sequences with  $H_{A1} \leq 7$  do all have a similar  $R_{ee}$  of roughly 23 Å. At  $H_{A1} = 8$ , this distance decreases slightly to 22 Å and falls abruptly between  $H_{A1} = 8$  and  $H_{A1} = 9$ . First, it reduces to 8 Å for sequence  $H_{A1} = 9$  and even further to 7 Å for target sequence N1. This abrupt change in the end-to-end distance represents the switch in the secondary structure.

In contrast to  $\langle\alpha\rangle$ ,  $\langle\beta\rangle$  and  $\langle R_{ee}\rangle$ , the curve following the average total energy,  $\langle E\rangle$ , does not display this abrupt change in structure along the pathway. Instead  $\langle E\rangle$  increases rather smoothly with increasing  $H_{A1}$  (figure 4d).

## 4.2 Validity of Reweighting for Different Observables

In the following, the results of the application of the reweighting algorithm are presented. Applying the algorithm to sequence A1, the reweighted averages of the four observables for the sequences along the mutational pathway towards N1 are predicted. The reweighted averages are shown in figure 4.

For the  $\alpha$ -content (figure 4a) there is a good agreement between the simulated and reweighted averages for those sequences with properties that resemble A1. This includes sequences with  $H_{A1} \leq 7$ . Also, the uncertainties in the reweighted averages are small for these sequences. The simulated averages are within the uncertainties until  $H_{A1} = 8$ . However, it can be seen that the reweighted average for  $H_{A1} = 8$  is associated with a larger uncertainty than the previous ones.

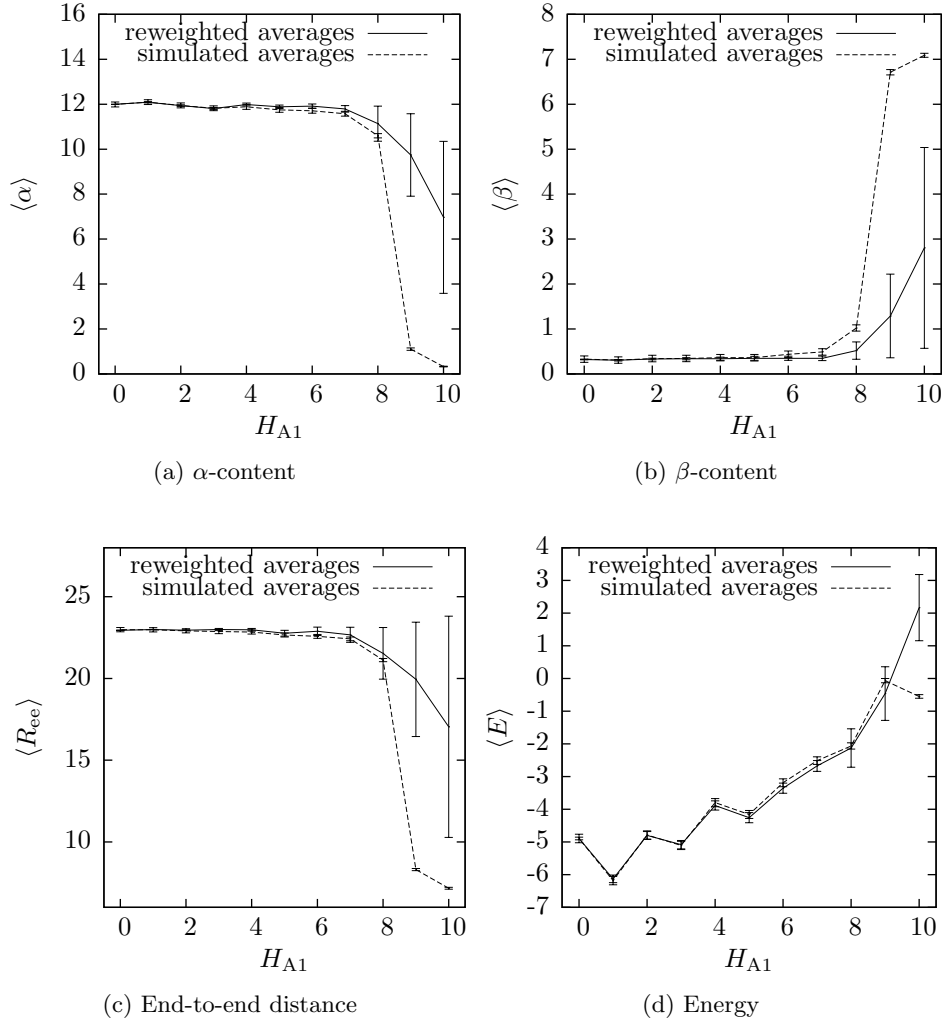


Figure 4: Comparison between the simulated and reweighted averages for all four observables that are considered in this project. On the top line, the evolution of the  $\alpha$ - and  $\beta$ -content towards N1 is shown. The bottom line presents the evolution of  $R_{ee}$  and the energy. The Hamming distances are given relative to sequence A1.

After the complete switch of the protein properties, the reweighted averages of the  $\alpha$ -content are far off from the simulated averages. Moreover, the reweighting method overestimates the  $\alpha$ -content for all sequences.

The reweighted averages of the  $\beta$ -content (figure 4b) fit the curve of the simulated evolution well up to Hamming distance  $H_{A1} = 7$ . The predictions deteriorate for the sequences afterwards. The reweighted averages lie generally below the simulated averages. Again, the line connecting the reweighted averages does not cross the line connecting the simulated averages. This time, the values of the reweighted averages are overestimated.

Many of the observations made for the  $\alpha$ -content can even be applied to the estimated end-to-end distance (figure 4c). Especially, all averages lie within the uncertainty of the reweighted averages for sequences with  $H_{A1} \leq 8$  and again the uncertainty increases abruptly from  $H_{A1} = 8$ . Just as for the  $\alpha$ -content, all reweighted end-to-end distances are overestimated.

Up to Hamming distance  $H_{A1} = 9$ , the reweighted energy averages are in very good agreement with the simulated averages (figure 4d). For the step from  $H_{A1} = 9$  to sequence  $H_{A1} = 10$ , a increase of the energy is predicted. This is in contrast to the simulations that have measured a slight decrease of  $\langle E \rangle$ . All simulated averages are within the uncertainties of the reweighted averages except for the predicted energy of N1. The uncertainty in the reweighted averages is small for those sequences with  $H_{A1} \leq 7$  and increases rapidly for the following sequences. This increase coincides with the switch in the native conformations of the sequences. For sequences with  $H_{A1} \geq 4$  the reweighting method tends to underestimate the conformational energies.

The same analysis can be performed in the reverse direction, i.e. by applying the reweighting algorithm to N1 and predicting the averages for the mutational pathway towards A1. These results are visualized in figure 5.

Overall, the reweighting algorithm works better in this direction. The estimated  $\alpha$ -content (figure 5a) is in good agreement with the simulated values. Especially, this is true for the Hamming distances  $H_{N1} = 1$ ,  $H_{N1} = 5$  and  $H_{N1} = 6$ . The sharp switch from  $\beta$ -sheet to  $\alpha$ -helix can be seen in the curve of the weighted averages but for  $H_{N1} = 2$  the simulated average is still outside the uncertainty of the estimate. For the sequences with  $1 < H_{N1} \leq 5$  the alpha content is slightly underestimated and for those sequences with  $H_{N1} \geq 7$  it is slightly overestimated. The uncertainty in the weighted averages increases with increasing Hamming distance.

For the  $\beta$ -content, there is a good correspondence between the reweighted and simulated averages (figure 5b). The simulated  $\beta$ -content lies outside the uncertainty of the reweighted averages for Hamming distance  $H_{N1} = 2$  only. As in the case of the  $\alpha$ -content, it can be seen that the two lines that follow the reweighted and simulated averages cross each other once. This time, it happens between the Hamming distances  $H_{N1} = 3$  and  $H_{N1} = 4$ . Generally, the uncertainties increase with increasing Hamming distance but their absolute values are smaller than those in the reweighted  $\alpha$ -content averages.

The predictions for the end-to-end distances (figure 5c) are good for Hamming distance  $H_{N1} = 1$  and sequences with  $H_{N1} \geq 4$ . In between these distances, the simulated averages lie outside the error estimate. With increasing Hamming distance, the reweighted averages approach the simulated averages again. This behavior could not be observed before. General trends are that  $R_{ee}$  is underestimated for all Hamming distances and that the uncertainty of the estimates tends to increase. This increase in the uncertainty, however, is not as obvious as for the uncertainties in the estimate of the  $\alpha$ -content.

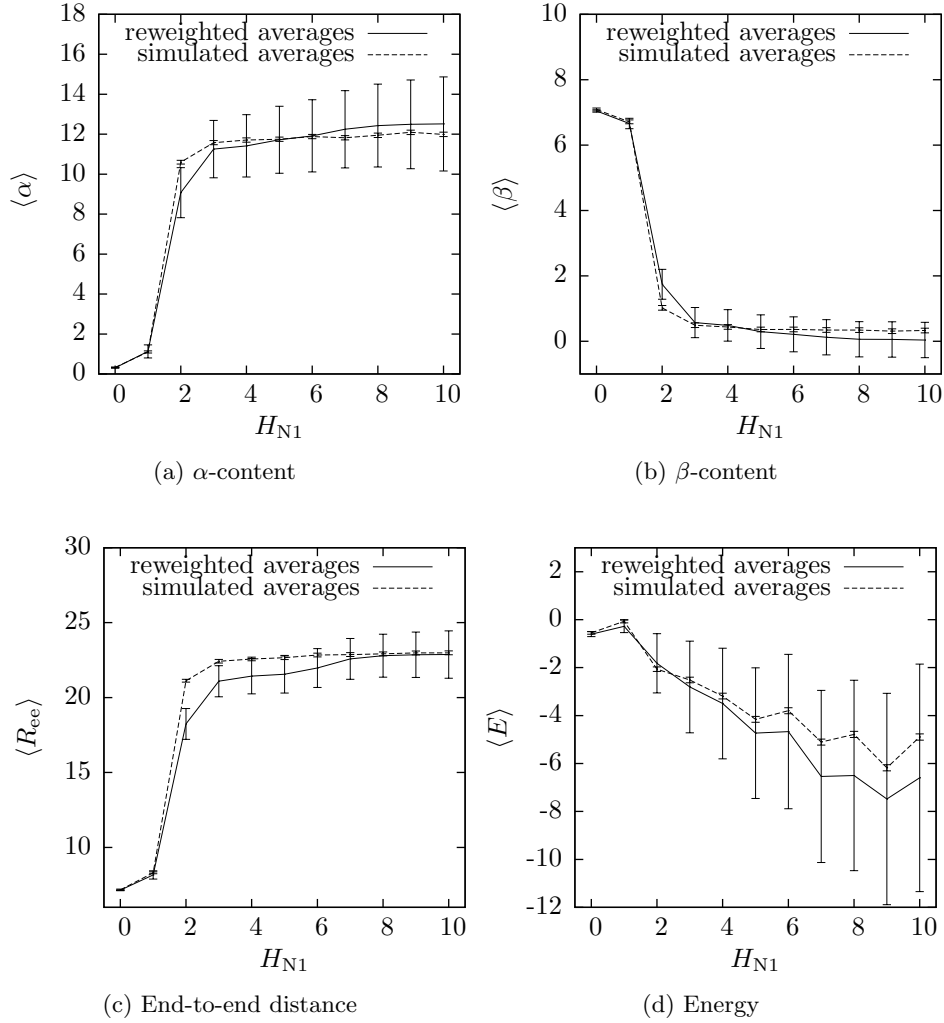


Figure 5: Comparison between the simulated and reweighted averages for all four observables that are considered in this project. On the top line, the evolution of the  $\alpha$ - and  $\beta$ -content towards A1 is shown. The bottom line presents the evolution of  $R_{ee}$  and the energy. The Hamming distances are given relative to sequence N1.

The reweighted energy averages (figure 5d) roughly reflect the general trend of an decreasing energy found in the simulations. The deviation between the weighted and simulated averages becomes bigger with increasing Hamming distance. It can be seen that all simulated averages lie within the uncertainty of the reweighted averages and that the reweighting algorithm underestimates the simulated energy averages for all Hamming distances. The uncertainties increase as the Hamming distance increases and are fairly big for sequences with  $H_{N1} \geq 2$ .

### 4.3 Bistable Sequence

The relation between Hamming distance and conformation space on one side and accuracy of the reweighted averages on the other side can be studied further. The conformation space of a bistable sequence contains both, sequences with a large  $\alpha$ - and  $\beta$ -content. Therefore, in theory, the reweighted averages based on the conformations of a bistable sequence should be accurate. In particular, it should be possible to predict the behavior around switches in the protein properties.

The primary structure of the bistable sequence as well as the averages for the relevant observables obtained after six simulations are shown in table 2.

Sequence	phppphphthhhphphph
$\langle E \rangle$	$0.923 \pm 0.03$
$\langle R_{ee} \rangle$	$13.0 \pm 0.08 \text{ \AA}$
$\langle \alpha \rangle$	$4.16 \pm 0.05$
$\langle \beta \rangle$	$4.60 \pm 0.03$

Table 2: The primary structure of the bistable sequence and the averages of the relevant observables with their associated uncertainties calculated after six simulations.

The Hamming distance between the bistable sequence and A1 is  $H = 8$ , whereas it is  $H = 2$  between the bistable sequence and N1. The average values of the four observables differ both from those measured for A1 and N1. The most interesting property of the bistable sequence is reflected in the averages of its  $\alpha$ - and  $\beta$ -content. On average, roughly the same number of amino acids are arranged in an  $\alpha$ -helix and a  $\beta$ -sheet. On average, the end-to-end distance of the bistable sequence assumes a value which is intermediate between those of A1 and N1 and the average energy of the bistable sequence is higher than those of A1 and N1.

The reweighted averages for all sequences from A1 to N1, based on the sampled conformations of the bistable sequence, can be seen in figure 6.

For the  $\alpha$ -content, the reweighted averages fit the simulated averages very well (figure 6a). All simulated averages lie within the uncertainties of the reweighted averages. The uncertainties increase slightly as sequence N1 is approached. Also, the reweighted averages do capture the switch from the mainly  $\alpha$ -dominated sequences to the  $\beta$ -dominated sequences. The  $\alpha$ -content of the sequences from A1 to P3 are underestimated.

For the  $\beta$ -content, there is a good agreement between the reweighted averages and the simulated averages. Again, all simulated averages lie within the uncertainties of the reweighted averages. The uncertainties are relatively constant over the  $\alpha$ -dominated sequences and increase for P1 and N1. Nevertheless, the switch from helix to sheet is captured. Between the sequences A1 and P3, the

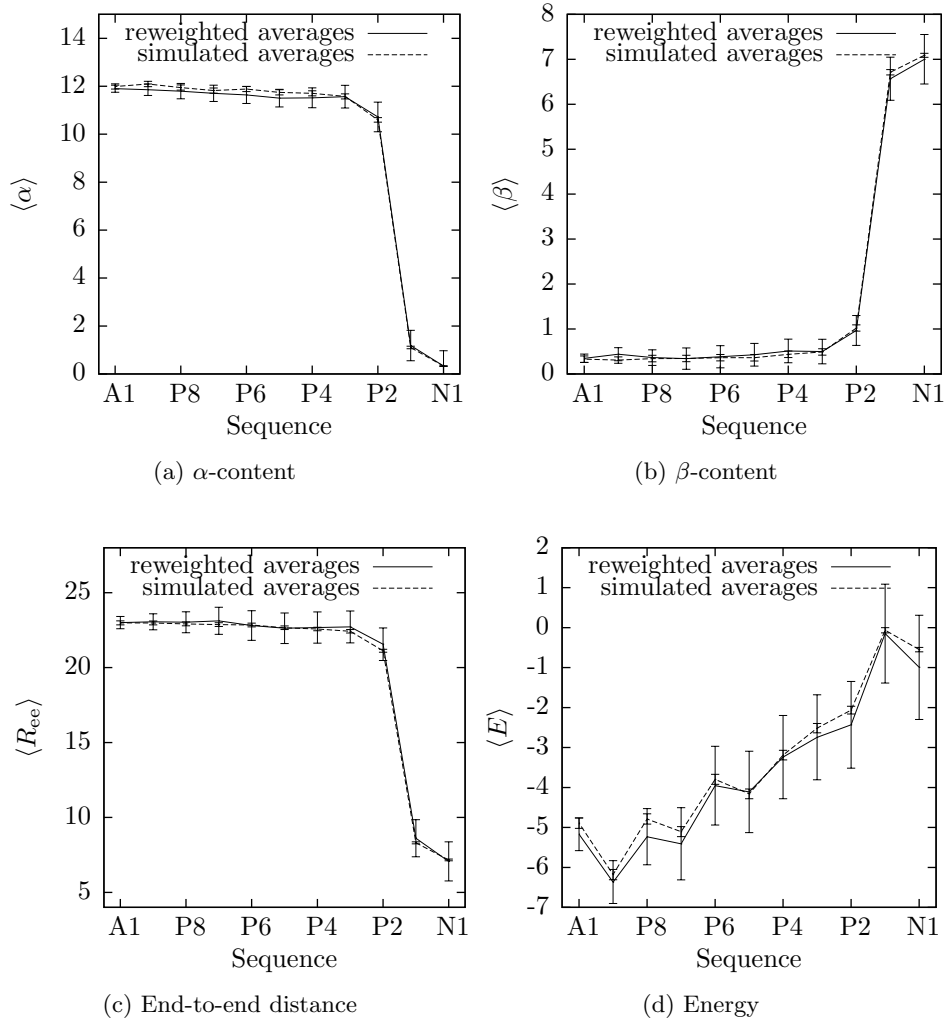


Figure 6: Comparison between the simulated and reweighted averages for all four observables based on the bistable sequence. On the top line, the evolution of the  $\alpha$ - and  $\beta$ -content is shown. The bottom line presents the evolution of  $R_{ee}$  and the energy. The x-axis is defined by the mutational pathway from A1 towards N1.

$\beta$ -content is overestimated and underestimated otherwise.

For the predicted end-to-end distances, basically the same analysis as for the  $\alpha$ -content applies (figure 6c). There is a very good agreement between the reweighted and the simulated averages and the simulated averages lie within the uncertainties of the reweighted averages. The switch in the protein properties can be read off from the graph. The uncertainties are relatively constant over the entire evolution. Unlike for  $\alpha$ -content,  $R_{ee}$  is overestimated for all sequences except for sequence P5.

The reweighted energy averages (figure 6d) follow the general trend from sequence A1 to N1, i.e. a increase in the conformational energy is predicted. All simulated averages lie within the uncertainties of the weighted averages. There is no tendency that the estimates deteriorate either towards A1 or N1 and the uncertainty is roughly constant for all sequences. The average energy is underestimated for all sequences.

#### 4.4 Neutral Net

After the promising results of the previous sections, the reweighting method can be tested on a concrete biophysical problem. In the studies of protein evolution it is often of interest to know the neutral nets that the sequences are part of. Therefore, in this section, the neutral nets which include sequences A1 and N1 are investigated.

The mapping of these nets is based on the conformations of sequences A1 and N1. This can be justified as the threshold values, that define the neutral nets, are chosen such that they are in close proximity of the  $\alpha$ - and  $\beta$ -contents of these sequences.

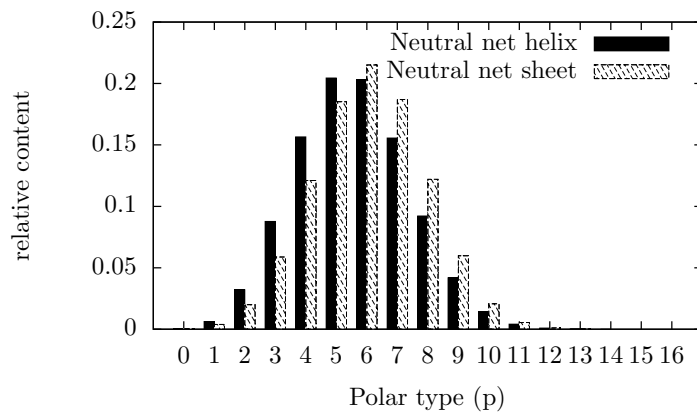
The neutral net of  $\alpha$ -helices is arbitrarily defined as being made up of those sequences that have a  $\alpha$ -content of at least 10.75. This value is roughly 0.90 times the  $\alpha$ -content of sequence A1. Correspondingly, the neutral net of the  $\beta$ -sheet consists of sequences with a minimum  $\beta$ -content of 6.40. The threshold of 0.90 times the respective  $\alpha$ - and  $\beta$ -contents of A1 and N1 was mainly chosen to keep the investigations manageable. Due to the enormous sequence space, the neutral nets can fast become very large which in turn requires long computation times. However, as the functionality of proteins is strongly dependent on their three-dimensional structure, it might still be realistic to assume such a low tolerance to changes in the native conformation.

According to one simulation, the neutral net of  $\alpha$ -helices defined in this way consists of 78,883 sequences. The corresponding neutral net of  $\beta$ -sheets is made up of 37,899 sequences. This means that the net of  $\alpha$ -helices is about twice as large as the net of  $\beta$ -sheets.

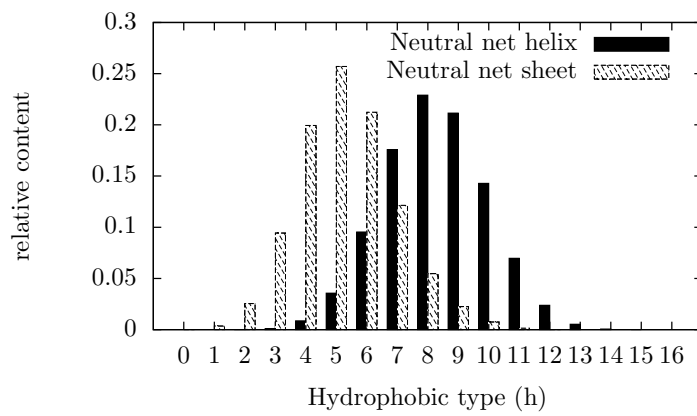
The typical compositions of the amino acids that are part of the two nets differ. The occurrences of proteins containing particular numbers of the different amino acids are shown in figure 7.

In the net of helices, the most frequently occurring numbers of polar amino acids in the sequences are five and six (figure 7a). The histogram over the number of polar amino acids in the neutral net is approximately symmetric around the most probable values. On average, the sequences in this net contain 6.0 polar amino acids. One sequence containing 14 polar residues has been identified as part of this net.

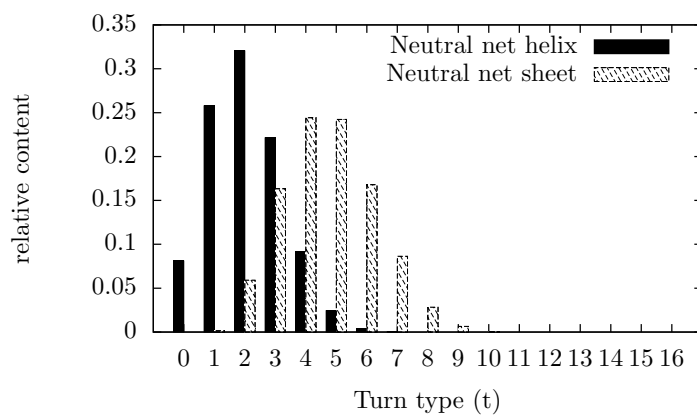
For the net of sheets, the most probable value of the polar residue content is 6. Further, the relative content of polar amino acids in this net is symmetrically



(a)



(b)



(c)

Figure 7: Histograms over the relative contents of sequences containing particular numbers of the respective types of amino acids.



around this top. The average sequence in this net contains 5.6 polar residues and there are three sequences that contain a maximum number of 13 of them.

The histograms that describe the relative content of hydrophobic residues in the two nets are shifted relative to each other (figure 7b). In the sheet net, the most probable number of h-type residues is five whereas it is eight in the net of helices. Neither of the distributions are symmetrical around their maximum values. Instead, in both cases, more sequences can be found on the right hand side of the maximum. This is even reflected in the averages. In the neutral net of sheets, the average sequence contains 5.3 h-type residues. The corresponding value for the helix net is 8.3. Moreover, no sequence identified as helix contains less than two hydrophobic residues and three of these sequences consist of 15 hydrophobic residues. In the sheet net the maximum number of hydrophobic residues found in a single sequence is 13.

The characteristic content of turn-type residues is very different in both nets (figure 7c). In the alpha net there is a fair number of sequences which do not contain any t-type amino acid. The distribution peaks at two turn-type residues and decays fast for t-type contents larger than two. There are no sequences with more than seven of these amino acids and the average content of it is 2.1.

In the net of sheets, the most probable values of t-type amino acid content are four and five residues. The distribution is not completely symmetrical around these values but there are more sequences to the right of the maxima. The average content of t-type amino acids in this net is 4.7. There are no sequences that do not contain any t-type amino acid and in the sequence containing most of them, there are ten of these residues incorporated.

The neutral nets are then analyzed further on the number of their direct neighbors. 87 direct neighbors could be found. Three pairs of them were randomly chosen to test the agreement between the predictions and simulations. The predicted values and simulated averages are shown in table 3.

Sequence	Net	Reweighted average	Simulated average
ththhhhhppphphhh	Helix	11.53	6.58
ththhhhhppptphhh	Sheet	6.80	2.89
tthhhhhphhphhhphp	Helix	11.61	10.81
tthhhhpthphhhphp	Sheet	6.49	2.85
tphtpppppphhpht	Helix	11.12	4.66
tphtptpppphhpht	Sheet	6.68	7.51

Table 3: The primary structure of the testing sequences and the nets they are identified to belong to. The reweighted and simulated averages show the  $\alpha$ -content for sequences that are identified as helices and the  $\beta$ -content for sequences that belong to the net of sheets.

As it can be seen, there is a poor agreement between the reweighted and simulated averages for almost all sequences. Only two out of the six tested sequences,

that were identified as either helix or sheet, do actually fulfill the requirements to be part of the respective nets.

## 5 Discussion

The results show, that in principle, the reweighting algorithm is capable of predicting the stability of native conformations. The test cases that are studied in this project, however, exhibit different behaviors.

In the first test case, sequence A1, a very stable  $\alpha$ -helix, was taken to predict the stability of the sequences that make up the mutational pathway towards sequence N1 which folds into a  $\beta$ -sheet. Generally, the reweighted averages that are calculated in this way, can be interpreted in terms of a critical Hamming distance,  $H_c$ . Predictions for sequences with Hamming distances  $H_{A1} \leq H_c$  are of reasonable accuracy. For sequences with Hamming distances  $H_{A1} > H_c$ , the reweighted averages are invalid. The critical Hamming distance is different for the different observables but lies in any case in the proximity of the switch in the secondary structure. Passing past  $H_c$  has a huge impact on the observables that describe the thermodynamic stability of the native conformations. Qualitatively, the reweighted averages of  $\langle\alpha\rangle$ ,  $\langle\beta\rangle$  and  $\langle R_{ee}\rangle$  indicate that the transition from helix to sheet is rather smooth. This is in contradiction to the simulated averages and potentially problematic when it is desired to map a neutral net. For such a task it is required that switches in the thermodynamic stability can be detected reliably and without delay. Another general trend is that the curves for the simulated averages and reweighted averages do not cross each other. The only point where this can be observed is in the energy reweighting between the Hamming distance  $H_{A1} = 9$  and  $H_{A1} = 10$ . Otherwise, the reweighting method does either over- or underestimate the simulated averages for all Hamming distances. This might indicate a systematic error which could be removed in a new, refined algorithm.

In comparison, reweighting, starting from sequence N1, displays different characteristics. Especially, there is no critical Hamming distance in the predictions of any of the investigated observables. This becomes clear when the results for the Hamming distances  $H_{N1} = 2$  and  $H_{N1} = 10$  are compared to each other. For  $H_{N1} = 2$ , none of the simulated averages  $\langle\alpha\rangle$ ,  $\langle\beta\rangle$ , and  $\langle R_{ee}\rangle$  lies within the uncertainties of the reweighted averages. On the other hand, for  $H_{N1} = 10$ , all of these averages can be predicted with some acceptable uncertainty. Despite the failure to predict these averages accurately for  $H_{N1} = 2$ , a clear drop in the  $\beta$ -content and a drastic increase for both the  $\alpha$ -content and  $R_{ee}$  can be noticed at this Hamming distance. Consequently, it can be concluded that the switch in the native conformation is captured at least to some extent and that sequence N1 might be a suitable starting point for mapping the neutral net of  $\beta$ -sheets. The relation between the curve of reweighted averages and simulated averages is much more complex for reweighting starting from N1. For the  $\alpha$ - and  $\beta$ -contents, the lines cross each other which means that the respective observables are underestimated for some Hamming distances and overestimated for others. For the end-to-end distance, the distance between the lines is bigger for intermediate Hamming distances compared to the largest distances and for the energy, the distances between the curves increase as  $H_{N1}$  increases. This richness in the behavior would certainly make it harder to implement any refinements as there is no simple offset that could make the reweighted averages more accurate. When it comes to the reweighted energy averages, it needs to be

mentioned that the uncertainty in most of them are so huge that the results can hardly be regarded as being suitable to make a statement on the actual energy content of any of the sequences with  $H_{N1} \geq 3$ .

Another important result emerges from the comparison between the capability to predict the  $\beta$ -content of N1 starting from A1 and the  $\alpha$ -content of A1 starting from N1. Even though the  $\alpha$ -content of N1 is roughly as low as the  $\beta$ -content of A1, the predictions of the respective content in the respective target sequence are qualitatively different. Starting from N1, the predictions for the stability of  $\alpha$ -dominated sequences are associated with a huge uncertainty but still acceptable. Starting from A1, however, gives poor estimates for the native conformation of  $\beta$ -dominated sequences. This observation can be related to the fact that sequence A1 is more stable than sequence N1 and therefore locked into a helix. Consequently, the conformation space that is visited during the simulations is small and in the reweighting, only very few conformations that do not have a big  $\alpha$ -content can be weighted up. On the other hand, the conformation space of sequence N1 contains more states that are different from a sheet as the average  $\beta$ -content is comparably low. This makes the reweighting algorithm more efficient as more averages can be weighted up and down.

The advantage of using a sequence whose structure is not locked can even be observed in the reweighting based on a bistable sequence, which gives very accurate results. The predictions are consistent with the simulated averages along the entire mutational pathway for all observables. Neither is there any critical Hamming distance nor is the capturing of the switch a problem. This means that the mapping of neutral nets, using the reweighting algorithm, should rely on the conformation spaces of bistable sequences.

Average reweighting using the the bistable sequence also indicates that not all observables can be predicted to the same accuracy. The uncertainties in the energy estimates is largest. This corresponds to the result of the reweighted energy averages using the conformations of sequence N1. These had to be rejected. The complications in predicting the energy might be due to the fact that the energy average is the sum of four individual energy contributions. In contrast, the parameters that describe the native conformations of the sequences are generally predictable to some better accuracy.

The results for the mapping of the neutral nets are not as convincing as those of the reweighting of the mutational pathway. Basically, the testing simulations show that the predictions for individual sequences are very poor. There might be different reasons for this. First of all, the mapping is based on the conformations of sequences A1 and N1. As discussed before, especially sequence A1 is unsuitable for the mapping of neutral nets. However, in the example, the threshold value of the  $\alpha$ -content is so close to the original value that, from the simulations of the mutational pathway, one could expect better predictions. For the testing sequences, that are identified to belong to the net of helices, with the poorest agreement between simulated and reweighted  $\alpha$ -content averages, these values deviate from each other by more than 50%. Such poor estimates were not to be expected after the results from the previous sections. Also, reweighting of N1 was not capable of mapping the neutral net of sheets even though this

sequence was capable of tracking the switch in the secondary structure on the mutational pathway towards A1. Another problem with the mapping is that it is based on the conformations of a single simulation of the respective sequences. Longer simulations of this sequence or several mappings of the nets followed by a comparison between the different outcomes would certainly contribute to make the mapping more precise.

Moreover, the composition of some sequences that are included in the neutral nets are very unexpected and unlikely to actually fold into the respective structures. For example, there are chains in the helix net containing 14 polar type amino acids. For such a sequence it should be energetically unfavorable to fold into a helix. The same applies to the sheet-identified sequences containing 13 h-type or p-type residues.

In spite of the shortcomings in the accuracy of the mapped nets, they still have some realistic features as well. The fact that there are many sequences without any t-type amino acids in the helix net is realistic. Further, no such sequence is found in the net of sheets. As the t-type acid is designed to be placed in the turns of sheets, this result is reasonable. Another realistic feature is the difference in the size of the nets, which is in agreement with the results of [11]. Hence, with some modifications, there might still be some potential in the reweighting algorithm to be used to map neutral nets.

## 6 Conclusions

In this thesis, it has been shown that the implemented reweighting algorithm, in principle, is capable of predicting the thermodynamic properties of short amino acid sequences in a simple protein model. The investigation of the relation between simulated and reweighted averages along a mutational pathway shows that the accuracy of the predictions are dependent on the properties of the sequence, which is used to compute the reweighted averages. The mappings of the neutral nets have shown that it is not enough to reweigh the averages from only one simulation to map out the net accurately. The huge discrepancy between the reweighted and simulated averages of test sequences also indicate that there might be complications in the reweighting method that were not encountered in the studies of the mutational pathway.

As the application of the algorithm is extremely time-saving compared to direct simulations, which in turn allows for studies that would be impossible otherwise, the implemented algorithm might still be a suitable starting point for further refinements, that can make the predictions more reliable.

## References

- [1] B. Alberts. *Molecular Biology of the Cell: Reference edition*. Number Bd. 1 in Molecular Biology of the Cell: Reference Edition. Garland Science, 2008.
- [2] C. H. Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.
- [3] B. A. Berg. Introduction to markov chain monte carlo simulations and their statistical analysis. *Markov Chain Monte Carlo. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap*, 7:1–52, 2005.
- [4] A. Bhattacharjee and S. Wallin. Coupled folding-binding in a hydrophobic/polar protein model: impact of synergistic folding and disordered flanks. *Biophysical journal*, 102(3):569–578, 2012.
- [5] E. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophysical journal*, 73(5):2393–2403, 1997.
- [6] H. S. Chan and E. Bornberg-Bauer. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences*, 96(19):10689–10694, 1999.
- [7] H. S. Chan and E. Bornberg-Bauer. Perspectives on protein evolution from simple exact models. *Applied bioinformatics*, 1(3):121–144, 2001.
- [8] T. E. Creighton. *Proteins: structures and molecular properties*. Macmillan, 1993.
- [9] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.
- [10] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.
- [11] C. Holzgräfe and S. Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophysical Journal*, 107(5):1217–1225, Sept. 2014.
- [12] R. C. King, P. Mulligan, and W. Stansfield. *A dictionary of genetics*. Oxford University Press, 2012.
- [13] D. P. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2009.
- [14] E. Marinari and G. Parisi. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.
- [15] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [16] P. Nelson. *Biological physics*. WH Freeman New York, 2004.
- [17] W. H. Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.