Interplay between Protein Thermodynamics and Protein Evolution

Bachelor Thesis

Andreas Buhr, March 15, 2006

Institut für Festkörperphysik
Technische Universität Darmstadt

AG Porto

Abstract

In this BSc thesis, correlations between the mutation process and the thermodynamical properties of a protein during evolution are determined using a simulation of a simple model of evolution, a three parameter mutation model and a structurally constrained model of protein folding. It is shown that a mutational bias influences protein stability, and that there exists an evolutionary pressure on this bias, which is important especially for small populations. Furthermore, the influence of different mutation rates onto the possibility of site-specific amino-acid distribution prediction has been tested.

Kurzfassung

In der vorliegenden BSc Arbeit wird der Zusammenhang zwischen den Eigenschaften des Mutationsprozesses und den thermodynamischen Eigenschaften des Proteins während der Evolution mittels einer Simulation analysiert. Ein einfaches Evolutionsmodell, ein Dreiparametermodell für die Mutationen sowie ein Modell der Proteinfaltung, welches auf dem Erhalt der Struktur basiert, wird zu Grunde gelegt. Es wird gezeigt, dass ein Hang zu A/T bzw. zu G/C im Mutationsmodell einen Einfluss auf die Stabilität des Proteins hat, und dass ein evolutionärer Druck auf diesen Hang existiert, der besonders in kleinen Populationen von Bedeutung ist. Weiterhin wird der Einfluss von verschiedenen Mutationsmodellen auf die Möglichkeit der Vorhersage von platzspezifischen Aminosäure-Verteilungen im Protein untersucht.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Recent work . . . 2

2 Concepts and Fundamentals **3**
- 2.1 Evolution and its Simulation . . . 3
- 2.2 DNA . . . 4
  - 2.2.1 Description . . . 4
  - 2.2.2 Codons . . . 5
  - 2.2.3 Genetic Code . . . 5
  - 2.2.4 Mutations . . . 6
- 2.3 Proteins . . . 7
  - 2.3.1 Amino Acids . . . 7
  - 2.3.2 Interaction . . . 8
  - 2.3.3 Hydrophobicity . . . 8
  - 2.3.4 Structure . . . 9
  - 2.3.5 Stability . . . 11
  - 2.3.6 The Fitness in Dependence on $E(C_{\text{nat}}, S)$ and α . . . 13
  - 2.3.7 The Principal Eigenvector . . . 15
  - 2.3.8 Prediction of Site-Specific Amino Acid Distributions . . . 15
- 2.4 Population Dynamics . . . 15
  - 2.4.1 Model for a Generation Step . . . 15
  - 2.4.2 Blind Ant Case . . . 16
  - 2.4.3 Analysis of Fixation Probabilities . . . 16

3 Results **21**
- 3.1 Description of the Simulation and Proteins Simulated . . . 21
- 3.2 Behavior of the System for Standard Conditions . . . 23
  - 3.2.1 Nucleotide Frequencies . . . 23
  - 3.2.2 Mutation Rates . . . 23
  - 3.2.3 Site-Specific Amino Acid Distributions . . . 24
  - 3.2.4 Site-Specific Mutation Rates . . . 24
  - 3.2.5 How Selection Works . . . 24
  - 3.2.6 Selection keeping up $|E|$ or α? . . . 26
- 3.3 Response of the System to Genetic Bias . . . 26

3.3.1 Overview . . . 26
3.3.2 On the DNA Level . . . 26
3.3.3 On the Amino Acid level . . . 28
3.3.4 On the Population Level . . . 30
3.3.5 Summary of Influence of Bias . . . 33
3.4 Response of a System to Change of Population Size . . . 33
3.4.1 Changes on the DNA Level . . . 33
3.4.2 On the Amino Acid Level . . . 34
3.4.3 On the Population Level . . . 34
3.4.4 Location in the (E, α)-Space . . . 35
3.4.5 Little Influence of Population Size . . . 35

4 Discussion and Outlook **37**

A Appendix **39**
A.1 Software Used . . . 39
A.2 Amino Acids . . . 39

List of Figures

2.1 Structure of Evolution 4
2.2 DNA 4
2.3 Example Codons 5
2.4 Mutation Probabilities 7
2.5 Amino Acid 7
2.6 Interaction and Hydrophobicity 8
2.7 Myoglobin 9
2.8 Distance Matrix of ATPE 10
2.9 Contact Matrix of ATPE 10
2.10 Energy Gap and Free Energy 12
2.11 Comparison of α_{thr} and α_{REM} 14
2.12 Fitness Function 14
2.13 Flow between States 17
2.14 Probability of Fixation 18
2.15 Probability of Fixation in Explicit Dynamics 19
2.16 Change in Fixation Probability for Fluctuation Population Size 19

3.1 Image and CM of ATPE 22
3.2 Image and CM of Lysozyme 22
3.3 Amino Acid Distribution in Lysozyme Site 12 and 14 24
3.4 Acceptance Probability P_{acc} vs PE Component 25
3.5 Where Lysozyme is in (E, α)-space 25
3.6 Content at 3^{rd} Codon Position vs AT-Bias 27
3.7 T Content at 2^{nd} Codon Position vs AT-Bias 27
3.8 'A' Content at Fist Codon Position 28
3.9 Site-Specific Mean Hydrophobicity vs AT-Bias for Lysozyme 28
3.10 Mean Hydrophobicity vs PE Component 29
3.11 Acceptance Ratio vs PE Component for Lysozyme 29
3.12 β_i vs PE Component for Lysozyme 30
3.13 Correlation of Exponential Parameter β_i and PE Component for Lysozyme 31
3.14 Fitness Distribution for different AT-Biases 31
3.15 Mean Fitness vs AT-Bias 32
3.16 Place in (E, α) Space 32
3.17 Dependence of DNA Frequencies on Population Size 33
3.18 Correlation of β_i and $c_i/\langle c \rangle$ in Dependence of Population Size 34

3.19 Mean Fitness for AT-Bias 0.1 and 0.9 vs Population Size 34
3.20 Difference in (E, α) Space 35

List of Tables

2.1 Genetic Code 5

3.1 Standard Conditions for a Population 23
3.2 Nucleotide Frequencies for Standard Conditions 23
3.3 Mutation Rates at the Different Codon Positions 24

1 Introduction

1.1 Motivation

In biological systems, proteins are of great importance. They are present everywhere in organisms and perform various tasks. They are essential for proper function of the cells. As enzymes, they perform chemical tasks and speed up reactions. As hormones, they are messengers and initiate processes, and in the cytoskeleton, proteins stabilize cells, to name just a few examples. However, the function and the design of proteins in not well understood.

A given native sequence of amino acids always folds into the same structure in nature. Even though this process of folding is determined by well-known physical laws, it is not possible to predict this structure when only the sequence is known. For a given protein, its function can often be determined, but usually, it is not possible to predict how the protein works in detail. Creating proteins ab initio, with a wanted function, is still beyond present day knowledge.

To further improve understanding of protein folding and function, what one can do is to look at proteins in nature, build models for them, try to explain what is seen, using the models, and improve them where their predictions fail.

Fortunately, lots of empirical data about proteins and their structure is available (for instance in the PDB, see A.1), so it is possible to create statistics about lots of proteins. Comparisons between predictions and real data is in many cases straightforward.

Due to the fact that many protein structures are known (often with a precision better than 1 Å), it is very promising to investigate different aspects of proteins in computer simulations. In this study is analysed, how protein sequences evolve in time under the constraint that the structure is kept fixed.

In this context, it is very important to keep in mind that proteins are encoded in DNA. When looking at the properties of proteins, some of the features seen are due to the requirements of stability and folding. Other effects are due to function, but there are also features or peculiarities, which can not be explained with the requirements of stability and functions; those features are due to the effect that a protein has to be robust against changes introduced by evolution. The situation is similar for evolution on the DNA level. Mutation processes in DNA cannot be understood without looking at what the DNA encodes. Properties of evolution, like fluctuating substitution rates, cannot be explained with simplified models, which do not take into account changes in fitness, like the model of neutral evolution [1].

In order to understand the system as a whole, numerical simulations of protein evolution, with constraints introduced by the stability of the protein combined with a simple model for mutations, are performed to investigate the interplay between mutations on

the DNA level on the one hand and of the structural requirements of proteins on the other. With the results of this simulation, is possible to better understand properties of both, proteins and evolution on the DNA level, which cannot be understood when focussing on only one part of the system.

For example in intracellular bacteria, a genetic bias has been observed, and also a tendency to hydrophobic proteins [2]. A better understanding of the dependency between these phenomena and some insight into reason for them can be obtained by this simulation.

1.2 Recent work

During the last years, various simulations of neutral evolution, extended with the requirements of protein thermodynamics have been performed by the group around Prof. Porto [3].

The SCN model (Structurally Constrained Neutral model of evolution [4]) is used to explain the non-Poissonian shape of occurrences of substitutions [5], as well as strong fluctuations and correlations in time of the substitution rate [6]; effects, which could not be explained by using a model of completely neutral evolution. Furthermore, high sequence dissimilarities in different proteins, sharing the same fold, are explained and highly conserved places in the protein are identified.

Simulations of evolution of proteins, which include genetic bias, but neither population dynamics nor non-neutral mutations, have been performed very recently [7].

2 Concepts and Fundamentals

2.1 Evolution and its Simulation

Before analysing details, it is important to see the structure of evolution.

Evolution is a process, which takes place on two layers at once. On the one hand, there is the DNA, which encodes all information, which are given to the next generation. DNA is a line through generations, it is handed from one generation to the next.

The driving force of evolution is mutation. A mutation is some change in the DNA that can be inherited to the next generation. It is not influenced by anything else in the system. Hence mutations can be discussed separately from the other parts of the system.

The other layer is the level of individuals in a population. Selection takes place on this level and does not directly care about changes in the DNA. The process of selection is only determined by the fitness of the included individuals. The fitness is a quantity which tells how successful the individual is in reproduction. For the evolution, only the number of its offspring is of interest. A fitness scale from 0 to 1 is used, but there is no meaning in the absolute value of these numbers. Only the relative fitness, i.e. the fitness of an individual in comparison to the individuals around it, matters. Selection, like mutation, can be investigated without taking into account other parts of the system (see section 2.4.3). So it is possible to look on selection in a population of individuals, having no properties but their fitness.

Between the two levels, DNA and mutation at the one hand, and the population and selection at the other hand, there has to be a link, some kind of genotype to phenotype mapping that forms a fitness on the basis of the DNA. In the simulation discussed in this thesis, this link is formed by the proteins. A protein is formed from the DNA, and the thermodynamic stability of the protein is calculated to determine its fitness which is then used for the selection process.

The DNA and its mutation on the one hand and the selection process on the other hand are, with the proteins as the link in between, the foundations of the simulation performed, and as such discussed in detail in this section.

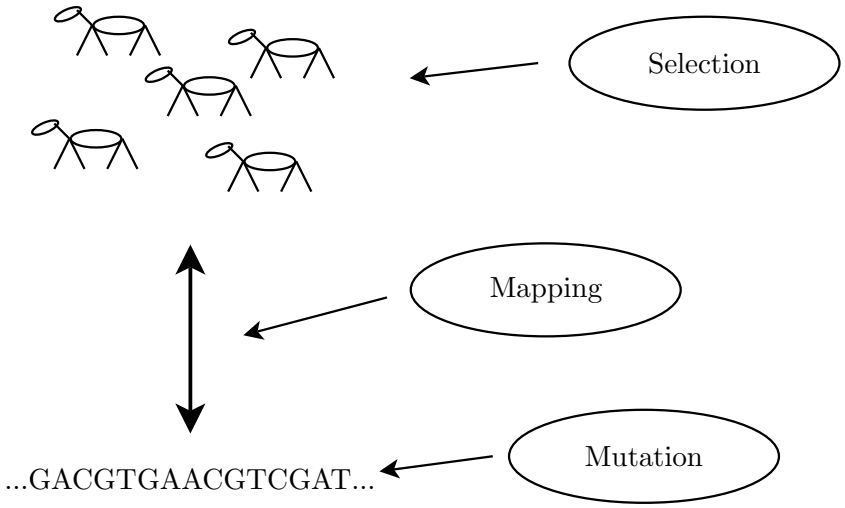

Figure 2.1: Structure of Evolution

2.2 DNA

2.2.1 Description

The DNA carries all the information about a creature to the next generation. It consists of a sequence of nucleotides: The two purines Adenine and Guanine and the two pyrimidines Cytosine and Thymine. These molecules form a double helix. The information stored in DNA is given by the sequence of these molecules. When talking about the DNA, the four molecules are usually abbreviated, using the four initials A,C,G, and T.

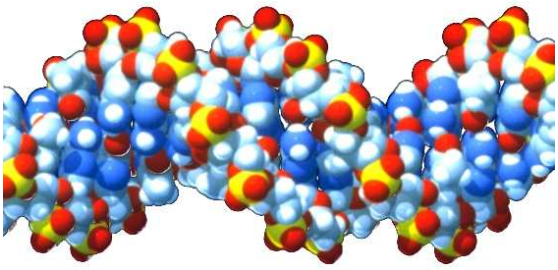

Figure 2.2: DNA (from Wikipedia)

Figure 2.3: Example Codons

code	aa	h	code	aa	h	code	aa	h	code	aa	h
TTT	F	0.41	TCT	S	-0.04	TAT	Y	0.32	TGT	C	0.27
TTC	F	0.41	TCC	S	-0.04	TAC	Y	0.32	TGC	C	0.27
TTA	L	0.42	TCA	S	-0.04	TAA	stop	-	TGA	stop	-
TTG	L	0.42	TCG	S	-0.04	TAG	stop	-	TGG	W	0.24
CTT	L	0.42	CCT	P	0.002	CAT	H	0.05	GGT	R	0.04
CTC	L	0.42	CCC	P	0.002	CAC	H	0.05	GGC	R	0.04
CTA	L	0.42	CCA	P	0.002	CAA	Q	0.03	GGA	R	0.04
CTG	L	0.42	CCG	P	0.002	CAG	Q	0.03	GGG	R	0.04
ATT	I	0.42	ACT	T	0.06	AAT	N	-0.03	AGT	S	-0.04
ATC	I	0.42	ACC	T	0.06	AAC	N	-0.03	AGC	S	-0.04
ATA	I	0.42	ACA	T	0.06	AAA	K	-0.01	AGA	R	0.04
ATG	M	0.17	ACG	T	0.06	AAG	K	-0.01	AGG	R	0.04
GTT	V	0.41	GCT	A	0.14	GAT	D	-0.12	GGT	G	-0.04
GTC	V	0.41	GCC	A	0.14	GAC	D	-0.12	GGC	G	-0.04
GTA	V	0.41	GCA	A	0.14	GAA	E	-0.04	GGA	G	-0.04
GTG	V	0.41	GCG	A	0.14	GAG	E	-0.04	GGG	G	-0.04

Table 2.1: Genetic Code, displaying both the coded amino acids as well as their hydrophobicity

2.2.2 Codons

The DNA sequence holds the information about the sequence of amino acids in proteins. When translating from DNA to the amino acid sequence, three nucleotides in the DNA encode for one out of 20 amino acids possible. These three nucleotides, which are responsible for one amino acid, are called a codon. So the DNA of coding regions can also be interpreted as a sequence of codons.

2.2.3 Genetic Code

When translating the DNA to a sequence of amino acids, nature uses a genetic code, which is common to most species. This code has to map 64 ($= 4^3$) different codons onto 20 amino acids, hence there have to be various synonymous codons. Which codons are synonymous and which codons belong to which amino acid is of great importance, because the structure of the genetic code influences the dynamics of the system.

The genetic code is given in table 2.1[8]: The hydrophobicity $h(a_i)$ is added here, which is discussed later in this section 2.3.3. The amino acids encoded are given in their corresponding one letter abbreviations.

The hydrophobicity is well structured in this table: Except for the codons beginning with TC, which code for Serine, a T in the first or second position is a sign for a very hydrophobic amino acid.

2.2.4 Mutations

As stated above, mutations are the driving force of evolution, so it is important to have close look at them. Mutations, occurring in nature, usually belong to one of these categories:

- Point mutations
  One nucleotide is exchanged with another one.
- Insertions:
  One or more nucleotides are added somewhere in the sequence.
- Deletions:
  One ore more nucleotides are removed.

In the simulation performed, only point mutations are considered. For those, the probabilities with which they occur are needed. Therefore a simple three-parameter model is used. The first parameter is the overall mutation probability. So, in the first approximation, all mutations have equal probability to occur. The next step is to introduce a higher probability for transitions. A transition is an exchange of a purine with another purine (A $\leftrightarrow$ G) or of a pyrimidine with another pyrimidine (C $\leftrightarrow$ T). All the other mutations are called transversions. In nature, there is observed that there are much more transitions than transversions. To meet this, the transition-transversion ratio (short: tt-ratio) is introduced. The third parameter accounts for the fact that the number of A and T is not always equal to the number of G and C. Such a genetic bias towards for instance A and T is modeled by giving all mutations starting from G and C a higher probability. To summarize: There is μ, the overall mutation probability, tt, the transition-transversion ratio, and b, the bias towards A and T.

The AT-bias used in the simulation is defined as follows: In the simulation $b = 0.5$ means that all mutations are equally probable, and for instance $b = 0.7$ means that mutation probabilities would lead to 70% AT without selection.

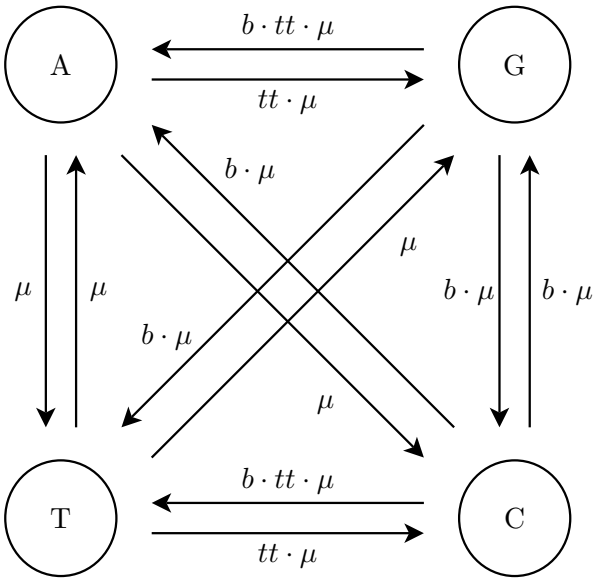

Figure 2.4: Mutation Probabilities between the four Nucleotides

2.3 Proteins

In the present context, the proteins are the link from the DNA to the fitness, so from the DNA a fitness has to be constructed.

2.3.1 Amino Acids

A protein is a sequence of amino acids, connected by peptide bonds and folded into its structure. Those amino acids are the building blocks of a protein. There are twenty

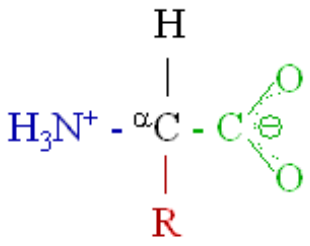

Figure 2.5: Amino Acid (from Wikipedia)

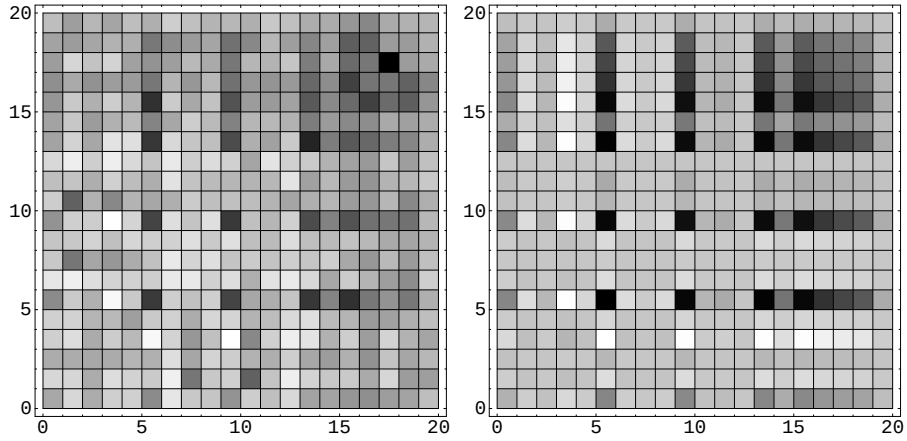

Figure 2.6: left: Interaction $I(a_i, a_j)$, right: Hydrophobicity $h(a_i)h(a_j)$

different amino acids (a detailed description is in Appendix A.2), which are characterized by the properties of their side chains. They can be hydrophobic, polar and/or charged, etc. These properties are essential for the structure of the protein. Amino acids, in contact with each other, interact and thereby stabilize the protein.

2.3.2 Interaction

There are a lot of effects to consider when evaluation the interaction of two amino acids in contact. For the simulations, this has to be simplified. A 20 x 20 interaction matrix with 210 numbers is used [9], that tells how much free energy is gained when to amino acids are in contact

$$I(a_i, a_j) \qquad a_i, a_j : \quad \text{amino acids} \tag{2.1}$$

2.3.3 Hydrophobicity

When looking at one amino acid, it is of interest how this amino acid interacts with others in general. Therefore the approximation

$$I(a_i, a_j) \approx h(a_i)h(a_j) \tag{2.2}$$

can be made where $h(a_i)$ and $h(a_j)$ are the hydrophobicities of the amino acids i and j. Into this hydrophobicity all effects of interaction are incorporated. The $h(a_i)$ values are obtained from the eigenvector to the largest (in absolute value) eigenvalue of the interaction matrix [10]. The similarity of $h(a_i)h(a_j)$ and $I(a_i, a_j)$ is clearly seen in figure 2.6, the correlation coefficient is 0.83 . Only the fine structure is lost. This hydrophobicity is well correlated with other experimentally obtained hydrophobicity scales [10].

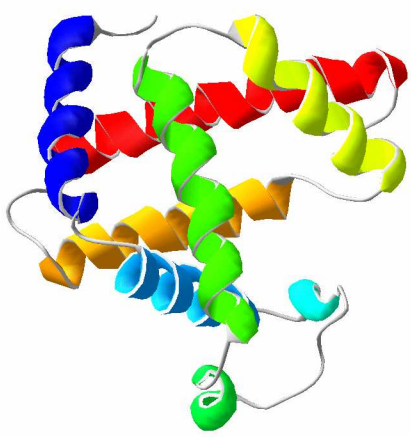

Figure 2.7: Myoglobin (from Wikipedia)

2.3.4 Structure

Primary, Secondary and Tertiary Structure

The structure of a protein is usually divided into three aspects:

1. The primary structure is the sequence of amino acids.
2. The secondary structure are patterns, like α-helices and β-sheets, where the sequence of amino acids forms a helix or two parts of the sequence lie next to each other.
3. The tertiary structure defines how the α-helices and β-sheets are arranged in space to form the folded structure.

The structure of a protein, or its fold, is highly conserved during evolution. Because of that, the structure is never changed throughout the simulation. The structure of a protein is usually determined using x-ray diffraction or NMR (the latter has less precision). So the structure is known in coordinates of all atoms of the protein. This is a lot of data, while most of the details are not of interest. Reduction of data is necessary. The first step in doing so is creating a distance matrix D, with the distances of all amino acids

$$D_{ij} = d(a_i, a_j) \tag{2.3}$$

In this representation, the information about all atoms is no longer contained, but can be reconstructed. When investigation protein stability, the distance is still more information than necessary. Of interest is usually just whether two amino acids are in contact, and thus stabilizing the protein, or not. Therefore the data is further reduced to a contact matrix C, where $C_{ij} = 1$ if amino acids i and j are in contact, and $C_{ij} = 0$ otherwise. Two amino acids are considered in contact, if the distance between their

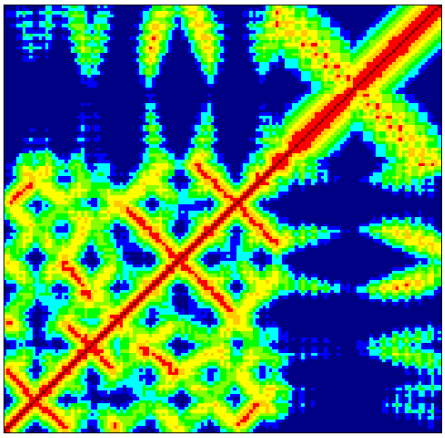

Figure 2.8: Distance Matrix of ATPE (created with WebMol from PDB data)

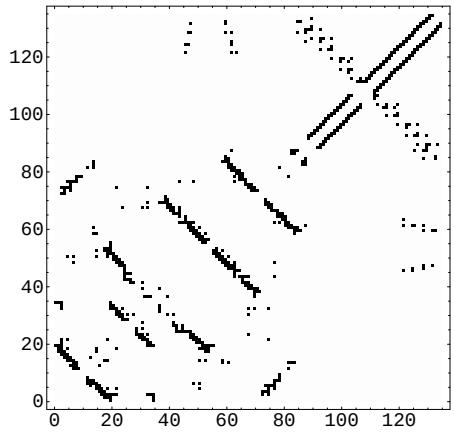

Figure 2.9: Contact Matrix of ATPE

closest non-hydrogen atoms is less than 4.5 Å. The figures 2.8 and 2.9 show the distance and contact matrix of ATPE, which is one of the proteins simulated, as an example. In the contact matrix, secondary structure, like β-sheets and α-helices, can be seen easily: Lots of contacts next to the diagonal are α-helices, diagonal or anti-diagonal stripes are β-sheets.

2.3.5 Stability

The protein is considered stable if it folds into its structure. There are two aspects of stability [3]:

- Stability against unfolding
  For a protein to fold properly, it is necessary that the free energy in the folded state is significantly below the free energy in the unfolded state. Otherwise the protein would not fold.
- Stability against misfolding
  The energy landscape has to be well correlated. That means, that there must not be some other structure than the native one, which has also a very low free energy. If there was such a structure, the protein would occasionally fold into this other structure and the folding process would be disturbed.

Calculation of the Stability against Unfolding

The free energy in a given structure has to be calculated. The structure is represented by a contact matrix C. The free energy by one contact is given by the interaction (see section 2.3.2). The energy of a sequence S in structure C is calculated as

$$E(S,C) = \sum_{i<j} C_{ij} I(a_i, a_j) \tag{2.4}$$

The absolute value $|E|$ must not lower significantly during evolution.

Calculation of the Stability against Misfolding

For all structures realized in nature, there is only one structure of low free energy[11]. Completely different structures have to have a much higher absolute free energy. Slightly different structures have to have a slightly higher absolute free energy. So first a measure for the similarity or dissimilarity of two structures is needed. Since these structures are represented by contact matrices, this measure has to be a function of two matrices.

One commonly uses the so-called contact overlap. The number of common contacts is divided by the number of all contacts.

$$q(C,C') = \frac{1}{N_c^*} \sum_{i<j} C_{ij} C'_{ij} \tag{2.5}$$

Where N_c^* is the maximum of the number of contacts in the two matrices.

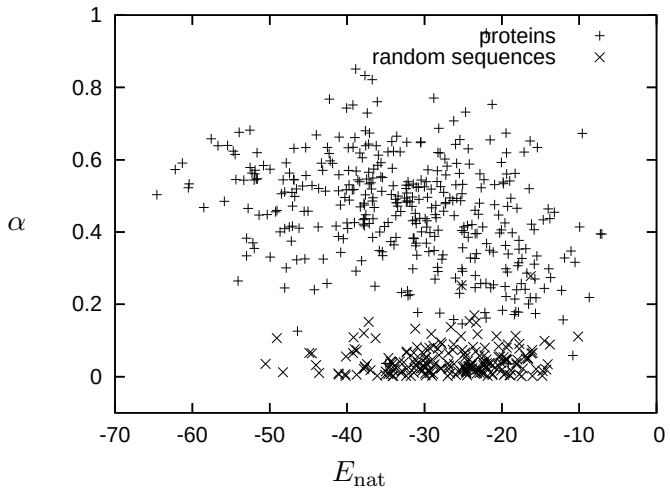

Figure 2.10: Normalized Energy Gap α against Free Energy E_{nat}

With this overlap it is possible to formulate the condition of stability:
For a protein sequence S with native structure C_{nat}, there has to be one number $\alpha(s) > 0$, with which for all other structures C the inequality

$$\frac{E(C_{\mathrm{nat}}, S) - E(C, S)}{E(C_{\mathrm{nat}}, S)} \geq \alpha(S)(1 - q(C, C')) \tag{2.6}$$

holds. I.e. the difference between the free energy in the native structure and in another structure, normalized by the free energy in the native structure, has to be at least the parameter α times the dissimilarity $(1 - q(C, C'))$.

Calculation of α by threading There is one very simple way to calculate an estimate for α. In (2.6) it is possible to calculate all values except of α. So α is obtained as

$$\alpha = \min\left(\frac{E(C_{\mathrm{nat}}, S) - E(C, S)}{E(C_{\mathrm{nat}}, S)} \frac{1}{1 - q(C, C')}\right) \tag{2.7}$$

This has to be done for all possible structures C. Unfortunately, it is not possible to just generate these structures. So a non-redundant subset of structures from the Protein Databank is taken and for all these structures (or substructures of these structures, if the proteins in the database are longer than the one in question) the right side of (2.7) is calculated. α is assigned the lowest value found.

This approximation works good. α calculated this way and E_{nat} as described in 2.3.5 for a set of 404 proteins of a length up to 200 amino acids (PDBSelect 25) and also for 200 random amino acid sequences are shown in figure 2.10. Nature optimizes α by selection. Random sequences do not have one good structure in the sense that α is large. Hence Eq. (2.7) allows to distinguish random sequences from native ones.

Calculation of α by REM

The calculation of α by threading works, but it is slow. It takes about 10 seconds on on a fast computer for one sequence. This is fine for some hundreds of proteins. But in a simulation, where of the order of 10^9 values of α have to be calculated, it is too slow. Furthermore, for large proteins, α is overestimated by threading, because there are not many alternative structures for them to compare with.

In the context of the random energy model (REM)[12] it is possible to estimate α very fast [13] by writing the free energy $E(C, S)$ for the alternative structures approximately as

$$E_{\mathrm{REM}}(S) \approx N_c\langle U\rangle - \sigma_U\sqrt{2N_c\log(m_N)} \tag{2.8}$$

Where N_c is the number of contacts, $\langle U\rangle$ is the average over all possible contact energies, σ_U is its standard deviation and m_N is the number of possible contact matrices for a protein of length N. This number is expected to grow exponentially with N, so it is assumed that

$$\log(m_N) \approx AN + B \tag{2.9}$$

The optimal values for A and B have been determined to [13]

$$A \approx 0.1 \qquad B \approx 4 \tag{2.10}$$

Putting (2.8) into (2.7) one gets:

$$\alpha_{\mathrm{REM}} = \frac{E(C_{\mathrm{nat}}, S) - N_c\langle U\rangle + \sigma_U\sqrt{2N_c\log(m_N)}}{E(C_{\mathrm{nat}}, S)(1-q_0)} \tag{2.11}$$

For q_0, the typical overlap of two unrelated structures, 0.1, is taken.

This α_{REM} is in good agreement with the α_{thr} generated by threading. These two α plotted against each other for the proteins in PDBSelect25 in figure 2.11. The correlation coefficient is $R = 0.69$.

2.3.6 The Fitness in Dependence on $E(C_{\mathrm{nat}}, S)$ and α

As long as $|E(C_{\mathrm{nat}}, S)|$ and α are above a certain threshold, the protein can fold, the fitness is one. When one of these parameters drops below the threshold, the protein is expected to cease to function, the fitness is zero.

In between, around the threshold, there is a smooth transition from one to zero.
For this transition, a Fermi distribution with $0.98E(C_{\mathrm{nat}}, S)$ and $0.98\alpha(C_{\mathrm{nat}}, S)$ as Fermi energy and $a \cdot E(C_{\mathrm{nat}}, S)$ and $a \cdot \alpha(C_{\mathrm{nat}}, S)$ for $k_{\mathrm{B}}T$ is used. a is a parameter to change the abruptness of the decay, it is set to $a = 0.02$ in this analysis.

$$E_{\mathrm{nat}} := |E(C_{\mathrm{nat}}, S)| \quad \alpha_{\mathrm{nat}} := \alpha(C_{\mathrm{nat}}, S) \tag{2.12}$$

The fitness is then

$$f = \frac{1}{e^{\frac{E_{\mathrm{nat}}\cdot 0.98 - E}{a\cdot E_{\mathrm{nat}}}} + e^{\frac{\alpha_{\mathrm{nat}}\cdot 0.98 - \alpha}{a\cdot \alpha_{\mathrm{nat}}}} + 1} \tag{2.13}$$

This dependency is shown in figure 2.12.

Note that other functional forms for the transition from 0 to 1 are also possible.

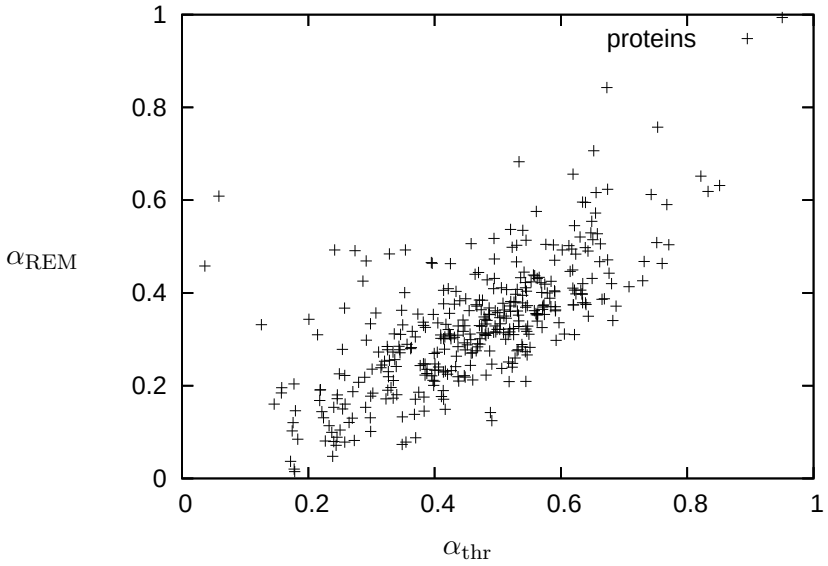

Figure 2.11: Comparison of α_{thr} and α_{REM}

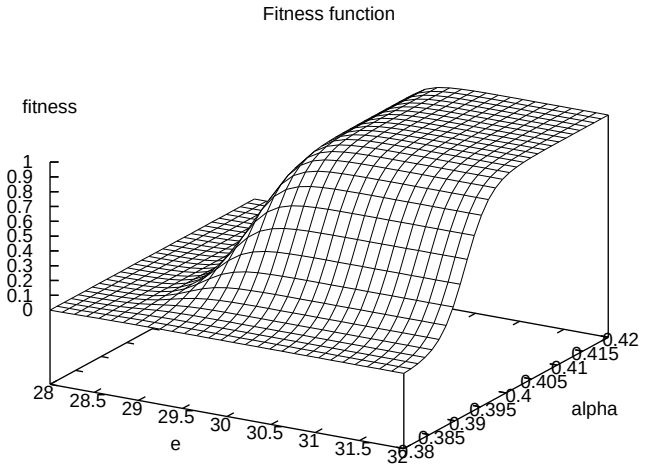

Figure 2.12: Fitness Function

2.3.7 The Principal Eigenvector

With these stability requirements, it is possible to give a statement where in the structure hydrophobic amino acids should be. At places, which are in contact with a lot of other hydrophobic amino acids, a hydrophobic amino acid should be, because this increases both $|E|$ and α. At places which are not in contact with a lot of hydrophobic amino acids, there should not be a hydrophobic amino acid. Putting a hydrophobic amino acid there does not hurt $|E|$, but the contacts in the CM are no longer optimized in comparison with random contacts, hence α gets down.

A simple measure for that is the number of contacts a place has: Places with a lot of contacts should be populated preferentially with very hydrophobic amino acids, places with few contacts preferentially with less hydrophobic amino acids. But this does not take into account that a hydrophobic amino acid is only advantageous if also the neighbouring are very hydrophobic. This is taken into account effectively using the principal eigenvector (PE) [14], the eigenvector to the largest eigenvalue of the contact matrix. A large component in the PE means that very hydrophobic amino acids are preferred at the corresponding place, a small component means that less hydrophobic amino acids are preferred [10].

2.3.8 Prediction of Site-Specific Amino Acid Distributions

Based on the description with the principal eigenvector, it is possible to predict, with which probability which amino acid is found in a specific place in the protein[15, 16]. The probability to find amino acid j in place i is an exponential function of its hydrophobicity

$$p_i(a_j) = \text{const} \cdot e^{-\beta_i h(a_j)} \tag{2.14}$$

The β_i values are site-specific and highly correlated with the PE [15, 16].

2.4 Population Dynamics

2.4.1 Model for a Generation Step

A population, like it exists in the real world, is very complex. With overlapping generations, fluctuating amount of offspring, fluctuating population size, etc.

To simulate it, it has to be simplified. In the first model, a population is a set of alleles. An allele is a group of individuals sharing the same genotype. From one generation to the next is a discrete step. In each step, the number of individuals in one allele scaled by some factors.

- For the whole population, there is not infinite space. If the population is too large, it will shrink. If the population is smaller than the available room, it will grow.

  With N as the population size, N_0 as the available space, the number of individuals is each population step is scaled by

$$\frac{N_0}{N} \tag{2.15}$$

- Better individuals are expected to increase in number at the expense of the worse individuals. So another scaling factor is the normalized fitness:

$$\frac{f}{\langle f \rangle} \tag{2.16}$$

Finally, the number of individuals in the next generation is randomized. It is Poissonian distributed with the expected mean. A maximum growth rate is introduced. Putting all this together yields

$$n_{\text{new}} = \text{Poisson}\left(\max\left(2, \frac{f}{\langle f \rangle}\frac{N_0}{N}\right) \cdot n_{\text{old}}\right) \tag{2.17}$$

Where n is the number of individuals in one allele.

2.4.2 Blind Ant Case: Simplify Simulation to Mutation-Fixation Steps

Let μ be the mutation rate, N the size of the population. There are on average $\mu \cdot N$ mutations per generation. If $\mu \cdot N \ll 1$, what is the case considered in the simulation, during most of the time no new mutants appear. Most of the time is dominated by the competition of the different alleles against each other. This ends in one allele becoming fixed, while all other alleles die out. After that, nothing happens until a new mutant appears. Thus, it is possible to see evolution as a sequence of discrete steps. In each step, a mutant appears. This mutant either dies out, or it takes over the whole population and is fixated.

Because of performance reasons, evolution is simulated in these discrete steps.
One important question that arises here is, how likely it is for a mutant to be fixated. The remaining part of this chapter discusses this probability.

2.4.3 Analysis of Fixation Probabilities

Theoretical Value for Fixation Probability

The model used for the generation step (i.e. the number of offspring is Poissonian distributed) is similar to the known Wright-Fisher Process. The main difference is, that this model has a fluctuation population size, while the Wright-Fisher Process has not.

For the Wright-Fisher Process, Sella and Hirsh [17] showed that the fixation probability is

$$\pi(i \to j) = \frac{1 - \left(\frac{f_i}{f_j}\right)^2}{1 - \left(\frac{f_i}{f_j}\right)^{2N}} \tag{2.18}$$

Fixation Probability calculated with Algebra

To verify this, a mathematical model for a population is applied. A population of size N can be in $N + 1$ states i. To be in state i means, that there are i individuals of mutant

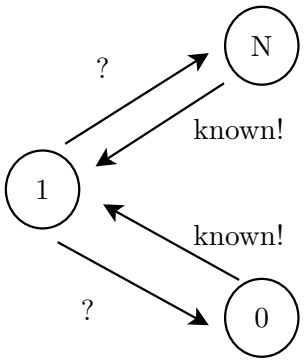

Figure 2.13: Flow between States of a Population

type. The probability to go from state i to state j is for neutral mutants

$$P(j = \text{Binomial}(N, i/N)) \tag{2.19}$$

With the fitness, it is

$$P(j = \text{Binomial}(N, \frac{fi}{N + i(f-1)})) \tag{2.20}$$

A generation step matrix S can be build with elements

$$S_{ji} = P(j = \text{Binomial}(N, \frac{fi}{N + i(f-1)})) \tag{2.21}$$

Of interest is the probability to be in state N at some time (i.e. fixation of the mutant), when it was in state $i = 1$ at the beginning. This can be solved in the following way: The generation step matrix is modified, so that $\frac{1}{1000}$ of the population, which is in state $i = 0$ or in state $i = N$, goes into the state $i = 1$ to avoid absorbing boundary conditions. The probabilities to be in state i in the equilibrium case can be calculated. By knowing the probability to be in states $i = 0$ or $i = N$, it is known how much flow there is from these two states to state $i = 1$. This is equal to the flow of this state to the states $i = 0$ and $i = N$. Therefore, the fixation probability is

$$P(\text{fixation}) = \frac{P(i = N)}{P(i = 0) + P(i = N)} \tag{2.22}$$

The probability to be in state i is the i-th component of the eigenvector to the largest eigenvalue of the modified generation step matrix.

This is in good agreement with formula (2.18). Some plots for different population sizes and the values of (2.18) are shown for comparison in figure 2.14.

Fixation Probability found in Explicit Population Simulations

To verify that the explicit population simulation is in agreement with these values, the simulation was run a large number of times, and it was observed how often a mutant

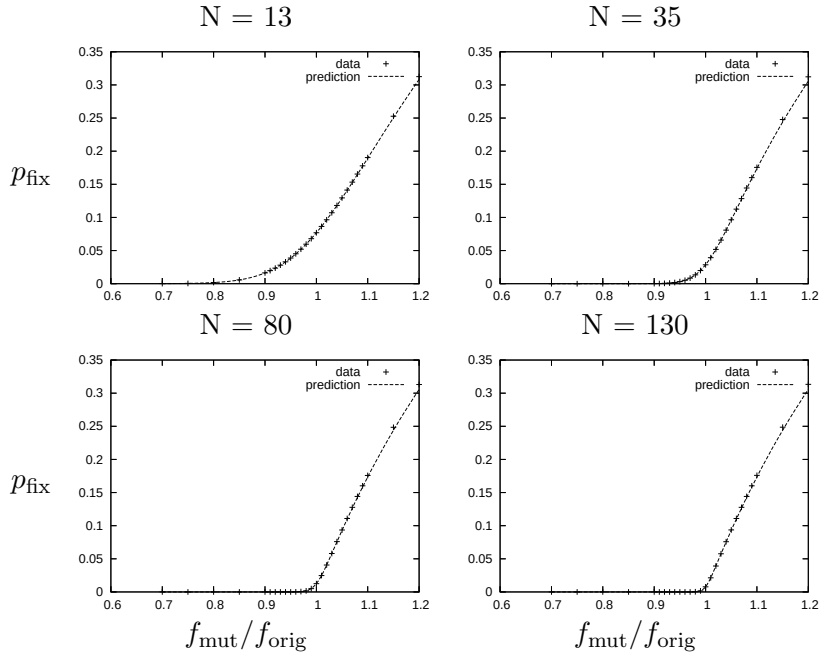

Figure 2.14: Probability of Fixation for different Population Sizes N

get fixated. It was run, until at least 1000 fixated mutants occurred, thus the error is about 3.2%. The results are plotted in figure 2.15. They are in good agreement. Only for small population sizes, one can see a little deviation. This is due to the reduction of the effective population size through fluctuations.

Reduction of Effective Population Size through Fluctuations

To estimate the probability of fixation in a fluctuating population, the expression is evaluated

$$\sum_{i=1}^{2N} P(\text{Poisson}(N) = i) \cdot P_i(\text{fixation}) \tag{2.23}$$

It is the sum over the probabilities, that a population of average size N has at the time of birth of the mutant i individuals times the probability of fixation in a population of size i. This is not correct, but it is a good approximation, because the critical part in the development is the first generation, and during this generation, the size of the population is Poisson-distributed.

The corrected fixation probabilities are plotted in figure 2.16. One sees that for population sizes as low as 30, there is nearly no effect. For a population size of 5 the fixation probability is increased by about 25%. For populations of a reasonable size, this effect is negligible. But it can explain the deviations seen in the fixation probability using explicit population dynamics for the population size of 13.

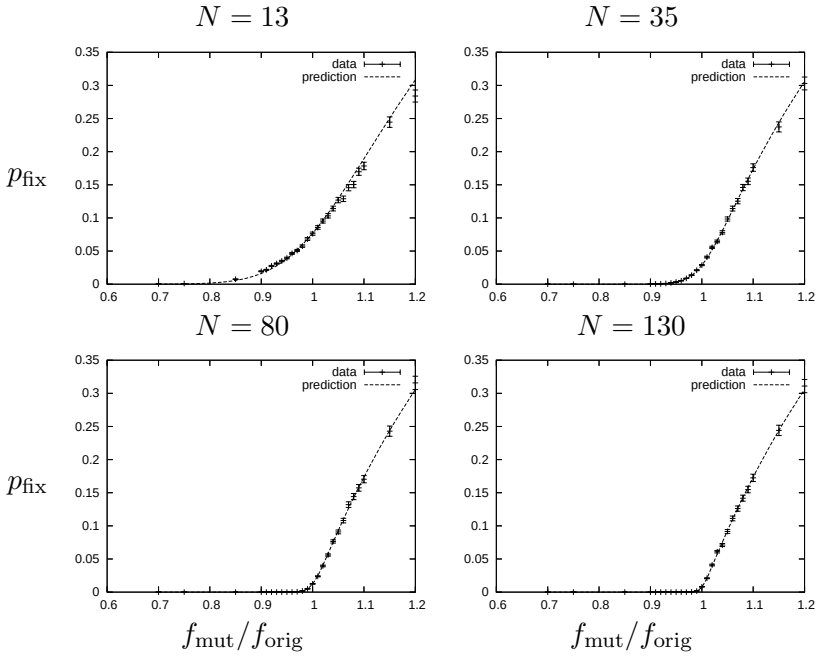

Figure 2.15: Probability of Fixation in Explicit Dynamics

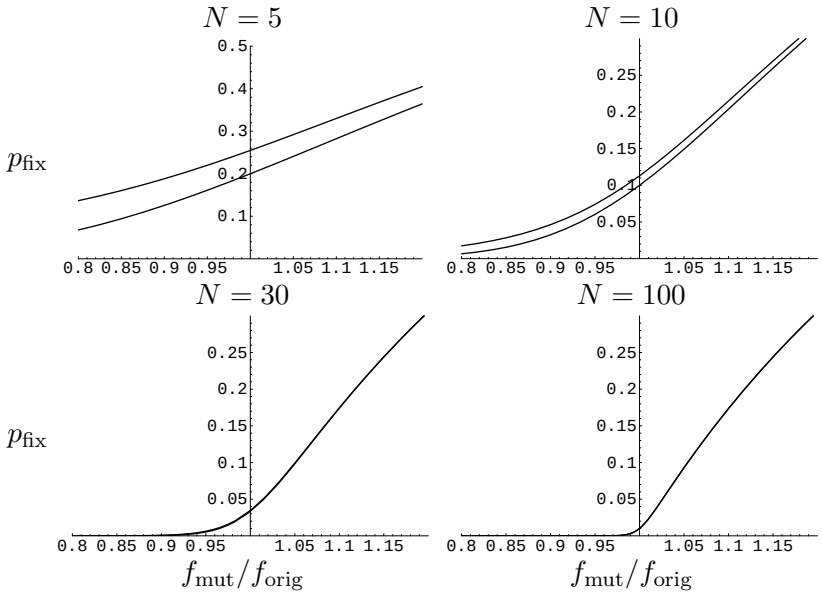

Figure 2.16: Change in Fixation Probability for Fluctuation Population Size

3 Results

In parallel to the structure of evolution and the structure of the simulation, there are three different views on the system.

- On the DNA level, the properties of DNA can be observed over time. Quantities of interest are for instance the nucleotide frequencies or substitution rates. It is important not to average over the whole DNA, but do statistics for each codon place, because the three places behave quite different.
- In the protein, site-specific analysis is possible. The probability to find an amino acid at a specific place should be an exponential function [15, 16]

$$P(a_j) \propto e^{-\beta_i h(a_j)} \tag{3.1}$$

  With β_i characteristic for each place in the protein. A small β_i ($\beta_i < 0$) accounts for places, where hydrophobic amino acids are preferred, which are in contact to a lot of other hydrophobic amino acids. $\beta_i > 0$ is expected at places which are connected to only few hydrophobic amino acids, so amino acids with low hydrophobicity are preferred there. These β_i are well correlated with the PE (see section 2.3.7) [15, 16].
- The third view on the system is to look on the population and the fitness. The evolutionary average of the fitness and also the distribution of the two stability parameters $|E|$ and α over time give valuable insight into the dynamics of the system, and into the process of selection.

3.1 Description of the Simulation and Proteins Simulated

A simulation of evolution of proteins is performed, in which a three parameter model for the mutations (the mutation probability, the transition-transversion ratio and the bias towards A and T) is used. The mutation probability drops out in the simulation, because the 'blind ant' case with $\mu N \ll 1$ is assumed. Each fixation at the population is treated as one step. The probability of fixation is dependent on the fitness of both the mutant and the other individuals in the population.

The two proteins simulated are ATPE, the ε subunit of ATP synthase (PDB id 1AQT) and Lysozyme (PDB id 3LZT). Both are small: 135 and 129 amino acids. While Lysozyme is single-domain, ATPE has two domains: The region with the β-sheets has only few contacts with the two α-helices. The prediction of size-specific amino acid distributions by use of the PE fail in the latter case, because the PE encodes only for the largest domain and the components for the smaller domain are much too small.

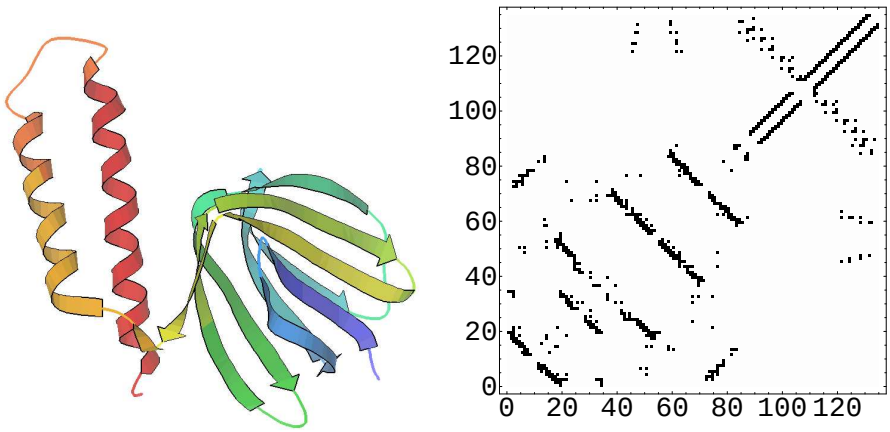

Figure 3.1: ATPE: image from PDB (created with KING), and Contact Matrix

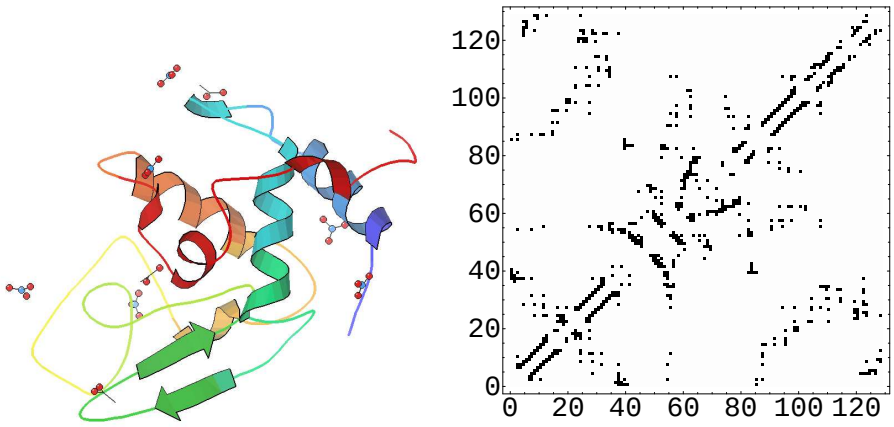

Figure 3.2: Lysozyme: image from PDB (created with KING), and Contact Matrix

3.2 Behavior of the System for Standard Conditions

It is worthwhile to have a close look on the behaviour of the system under standard conditions, before looking at dynamics when changing parameters. The standard conditions are chosen as given in table 3.1.

population size	10
AT bias	0.5
tt ratio	2

Table 3.1: Standard Conditions for a Population

3.2.1 Nucleotide Frequencies

For an AT bias of 0.5, the same frequencies for all four nucleotides A,C,G, and T are expected. However, different frequencies are observed (see table 3.2). The reasons for these deviations become obvious when looking at the effect of a changed AT bias later on.

3.2.2 Mutation Rates

It is very interesting to look which place in the codons is mutated: For both ATPE and Lysozyme, most mutations are at the third place. These mutations are most often synonymous (see the genetic code 2.2.3). Mutations in the first or second place are most often deleterious.

ATPE:			
	N_1	N_2	N_3
A	26 %	22 %	23 %
C	26 %	26 %	26 %
G	28 %	24 %	24 %
T	20 %	28 %	26 %

Lysozyme:			
	N_1	N_2	N_3
A	26 %	22 %	23 %
C	26 %	28 %	26 %
G	27 %	24 %	24 %
T	20 %	26 %	26 %

Table 3.2: Nucleotide Frequencies for Standard Conditions

	ATPE	Lysozyme
1^{st}	34.1 %	33.7 %
2^{nd}	25.7 %	27.4 %
3^{rd}	40.1 %	38.9 %

Table 3.3: Mutation Rates at the Different Codon Positions

3.2.3 Site-Specific Amino Acid Distributions

For two places in the Lysozyme protein, one inside the structure (PE component $c_i/\langle c\rangle > 2$) and one at the surface (PE component $c_i/\langle c\rangle < 0.5$) the probability to find amino acid j is plotted against its hydrophobicity in figure 3.3. The predicted exponential distribution (2.3.8) can be seen. A fit with an exponential function $e^{-\beta_i h(a_j)}$ is included.

3.2.4 Site-Specific Mutation Rates

Sites inside the protein with a lot of contacts (large PE component) and sites at the surface, with few contacts (low PE component) are highly conserved. Sites in between are more tolerant to mutations (see figure 3.4).

3.2.5 How Selection Works

Most mutation lower either $|E|$ or α or both, hence these parameters have to be kept up by selection. During evolution, most of the time the protein is somewhere in (E, α) space with both components only slightly larger than the threshold, see figure 3.5.

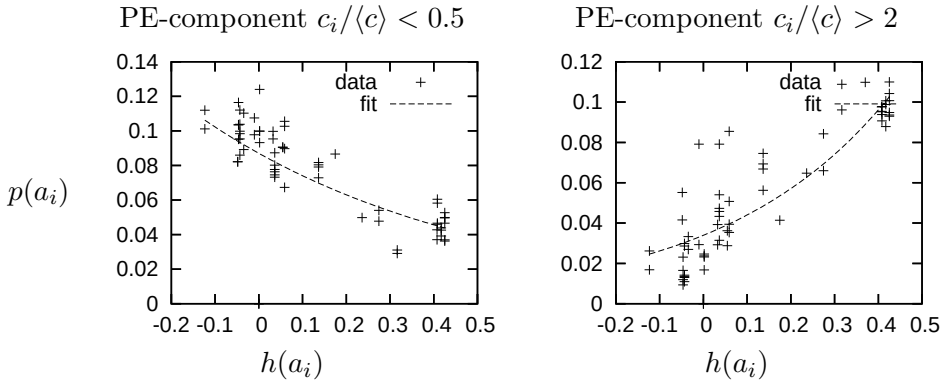

Figure 3.3: Amino Acid Distribution in Lysozyme Site 12 and 14

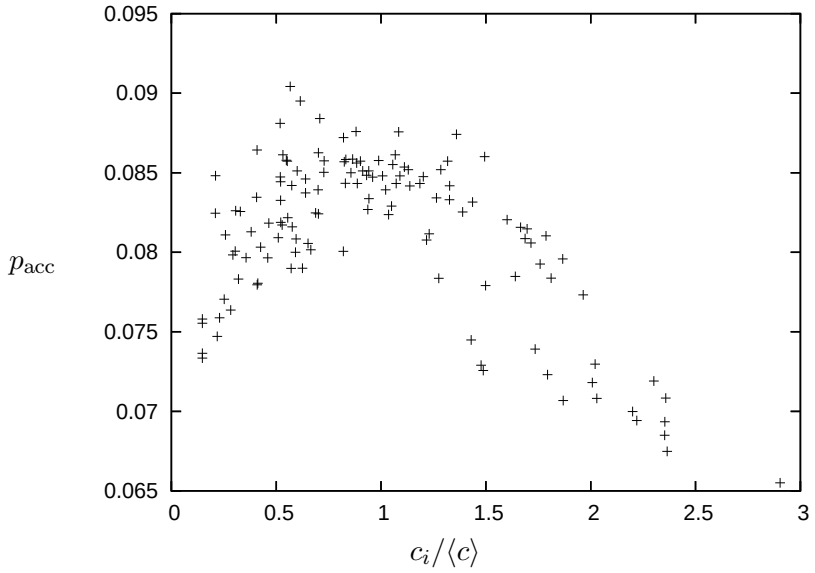

Figure 3.4: Acceptance Probability P_{acc} vs PE Component

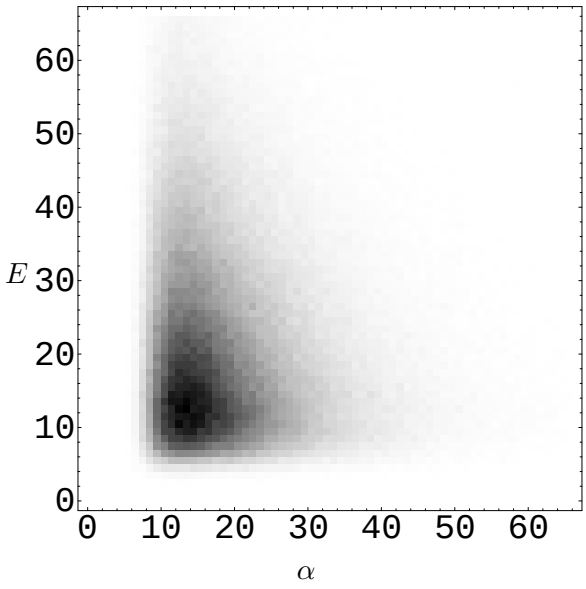

Figure 3.5: Where Lysozyme is in (E, α)-space. The distance to the threshold is given in units of $\alpha_{\mathrm{thr}}/100$ and $|E_{\mathrm{thr}}|/100$

3.2.6 Selection keeping up $|E|$ or α?

When

$$\frac{E}{E_{\text{thr}}} > \frac{\alpha}{\alpha_{\text{thr}}} \tag{3.2}$$

the population is closer to the α-threshold than to the E-threshold. The fraction of time during evolution, while this holds, is called w. A large ($w \approx 1$) value means that selection has to care about keeping up α most of the time. A low ($w \approx 0$) value accounts for more problems with E than with α. For the standard conditions

$$w = 0.59 \quad \text{ATPE} \tag{3.3}$$

$$w = 0.66 \quad \text{Lysozyme} \tag{3.4}$$

So both parameters have to be kept up by selection under standard conditions.

3.3 Response of the System to Genetic Bias

It is of great interest how the system reacts on a genetic bias. It has to be remarked, that in nature, not necessarily the remaining parts of the system just react on a bias, but also the bias can react on changes on the population level. In this simulation, the bias towards AT is fixed, hence the remaining parts of the system react. Biases from 1% AT-content to 99 % AT-content are tested. Again, the analysis is divided into three parts. All other parameters are held at the standard value.

3.3.1 Overview

A bias towards AT causes more hydrophobic amino acids (see section 2.2.3). More hydrophobic amino acid cause a stable (large) $|E|$, but an unstable (small) α.

3.3.2 On the DNA Level

Small Selection Pressure on Codon Position 3

At the third codon position, the observed content is nearly equal to the value demanded by the genetic bias. This is due to the fact that the 3$^{\text{rd}}$ codon position is usually synonymous, so there is almost no selection pressure on it. For an AT bias > 0.5, a small deviation is seen (figure 3.6). The cause is the genetic code: When the first two nucleotides are both A or T, the third nucleotide is not synonymous. For the first two nucleotides being both G or C, the third place is synonymous.

Conservation of T Content at the Second Codon Position

Content of T at the second codon position, c_{T2}, is stabilized in the system. There is a native c_{T2}, specific for the protein, which changes only slightly when the system is perturbated by a genetic bias. The cause of this behaviour is the structure of the genetic code. A change to or from T in the second position is almost always a change in the hydrophobicity of the encoded amino acid.

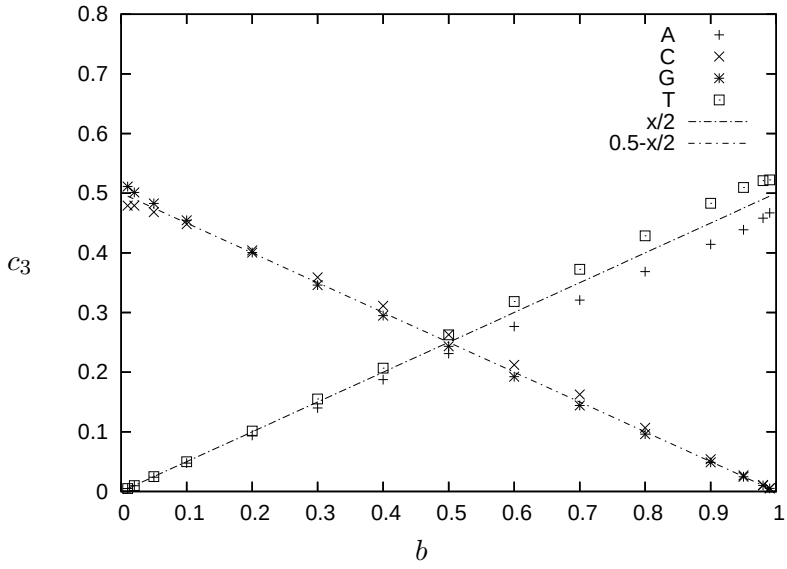

Figure 3.6: Content at 3rd Codon Position vs AT-Bias

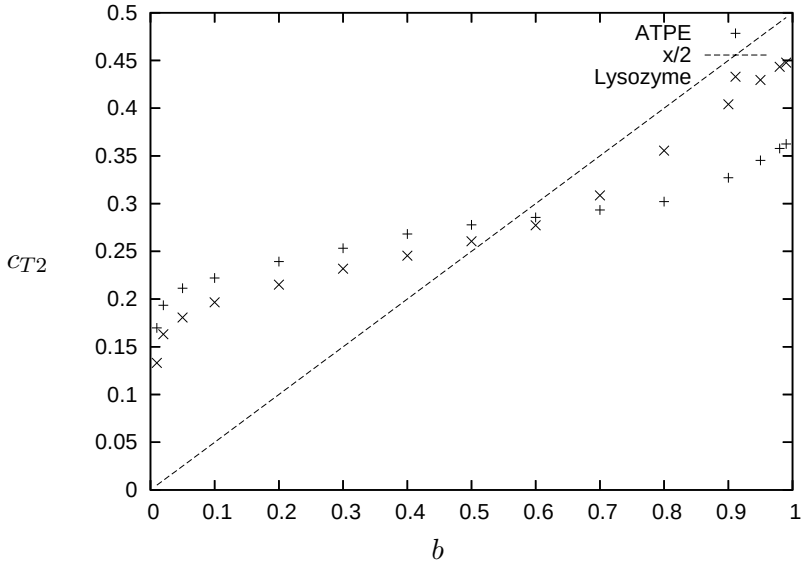

Figure 3.7: T Content at 2nd Codon Position vs AT-Bias

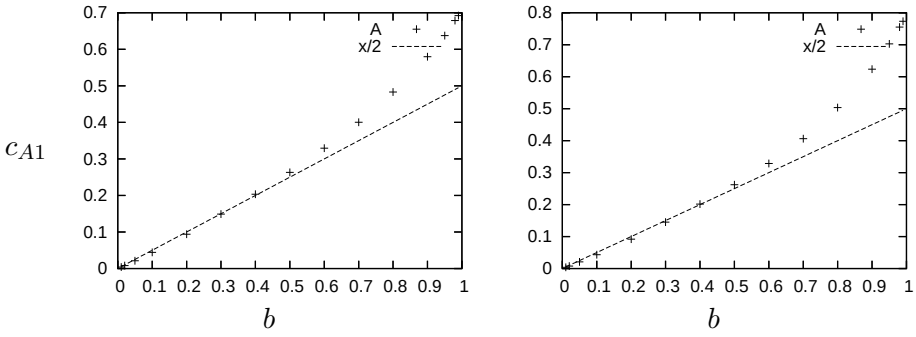

Figure 3.8: 'A' Content at Fist Codon Position vs AT-Bias, left: Lysozyme, right: ATPE

Overreaction of A content at the 1^{st} Codon Position

The content of A at the first and second codon position, c_{A1} and c_{A2}, are not stabilized like c_{T2}. In contrast, its reaction to genetic bias is stronger than expected. Because mutations to T are hindered, the system resorts to mutations to A for a large bias b.

3.3.3 On the Amino Acid level

Change of Mean Hydrophobicity

Without an AT bias ($b = 0.5$), the site-specific mean hydrophobicity $\langle h \rangle$ is highly correlated with the PE component c_i. For sites with high c_i, a low b accounts for too hydrophilic amino acids in this place. For sites with low c_i, a high b causes too hydrophobic amino acids in that place. This correlation is gets better for a bias towards GC, and worse for a bias towards AT (figure 3.10). An AT-bias disturbs this dependency.

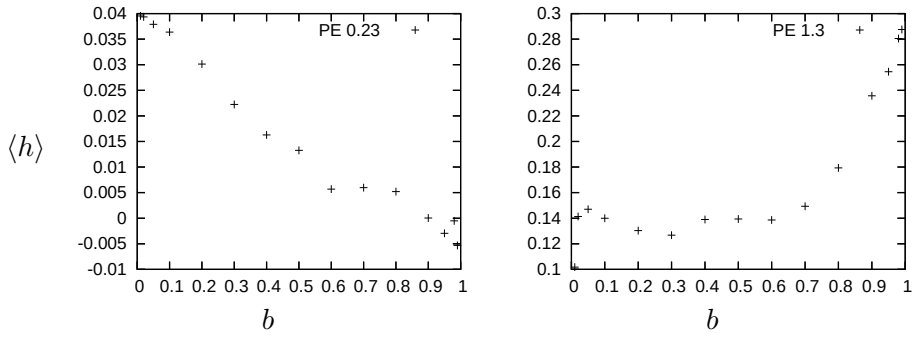

Figure 3.9: Site-Specific Mean Hydrophobicity vs AT-Bias for Lysozyme

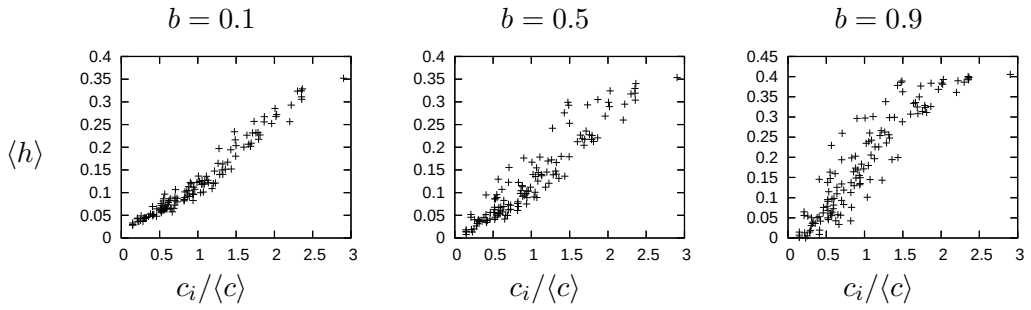

Figure 3.10: Mean Hydrophobicity vs PE Component for Lysozyme and different AT-Biases

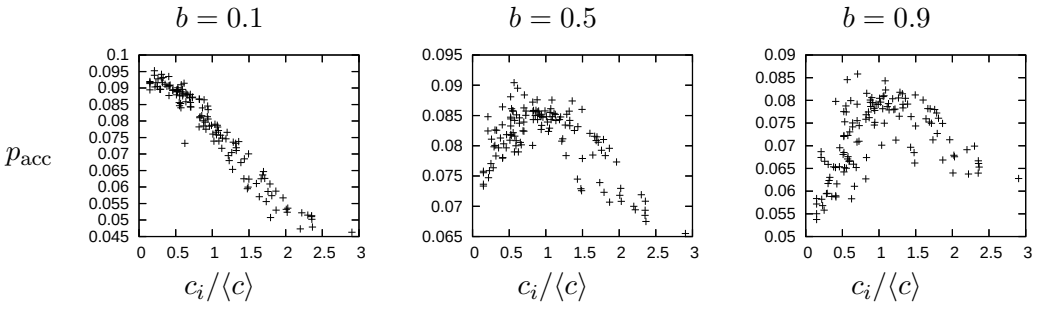

Figure 3.11: Acceptance Ratio vs PE Component for Lysozyme

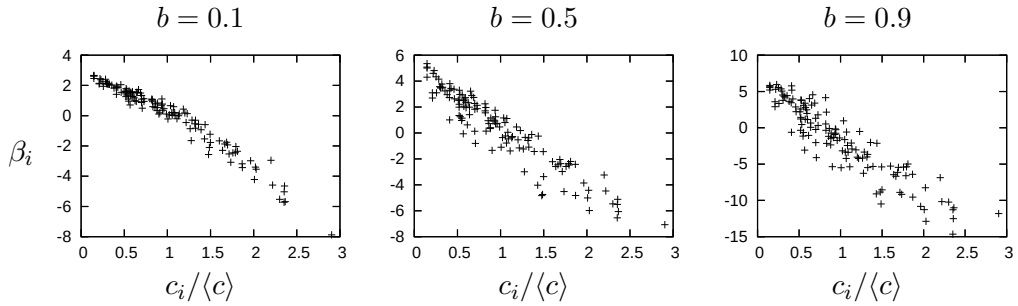

Figure 3.12: β_i vs PE Component for Lysozyme

Site-Specific Acceptance Ratios

Where mutations are accepted depends on whether the amino acids the genetic code yields (more hydrophobic or more hydrophilic amino acids) is in line with what is preferred by the protein at this place.

Figure 3.11 shows that: The tolerant places ($c_i/\langle c \rangle \approx 1$) are mutated most often. For $b = 0.1$ most mutations are towards less hydrophobic amino acids, these are accepted at places which are not responsible for holding together the protein. Mutations to very hydrophobic amino acids occur most often for a large b. b being 0.9, the places with a lot of contacts deep inside the protein ($c_i/\langle c \rangle > 1.5$) are mutated most often.

Site-Specific Amino Acid Distributions

Without any influence of genetic code and bias, the probability to see an amino acid is an exponential function of its hydrophobicity. A bias towards GC stabilizes this behaviour, while a bias towards AT disturbs it, and a more general ansatz is needed.

The stabilizing behaviour of an GC bias is seen in figure 3.13. The correlation of the exponential parameters β_i and the corresponding components of the principal eigenvector is best for a bias towards GC. For $b < 0.1$, the correlation drops to less than 0.9.

An extreme genetic bias ($b < 0.1$ or $b > 0.9$) disturbs this dependency, because some of the amino acids gain a higher priority. Especially places with only a small dependence β_i are very sensitive.

3.3.4 On the Population Level

In a population, the fitness is the only property of the proteins. The distribution of the fitness throughout evolution is shown in figure 3.14.

Most of the time, fitness stays up very close to one. Only 3 % of the time the population is below a fitness of 0.9 for $b = 0.5$. For $b = 0.1$ and $b = 0.9$, 8% of the time the population is below a fitness of 0.9.

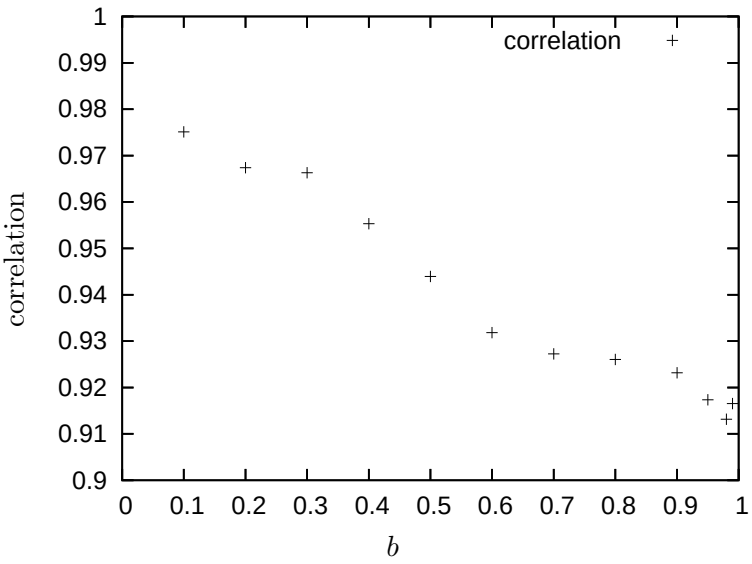

Figure 3.13: Correlation of Exponential Parameter β_i and PE Component for Lysozyme

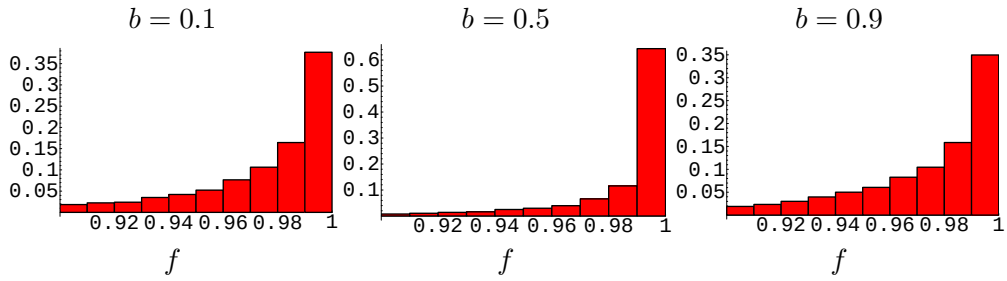

Figure 3.14: Fitness Distribution for different AT-Biases

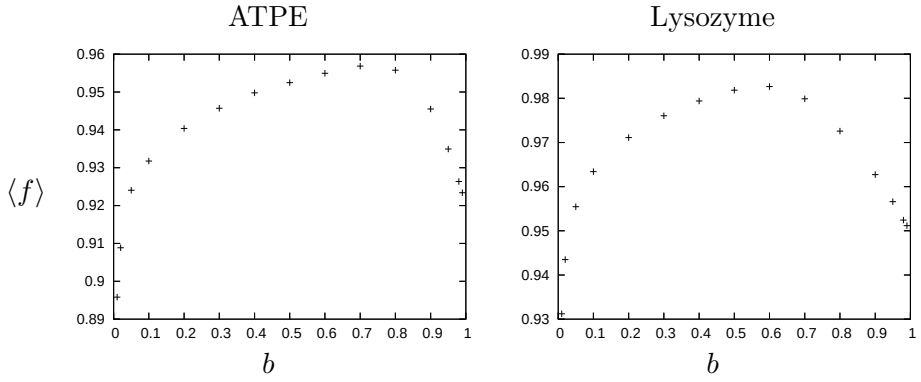

Figure 3.15: Mean Fitness vs AT-Bias

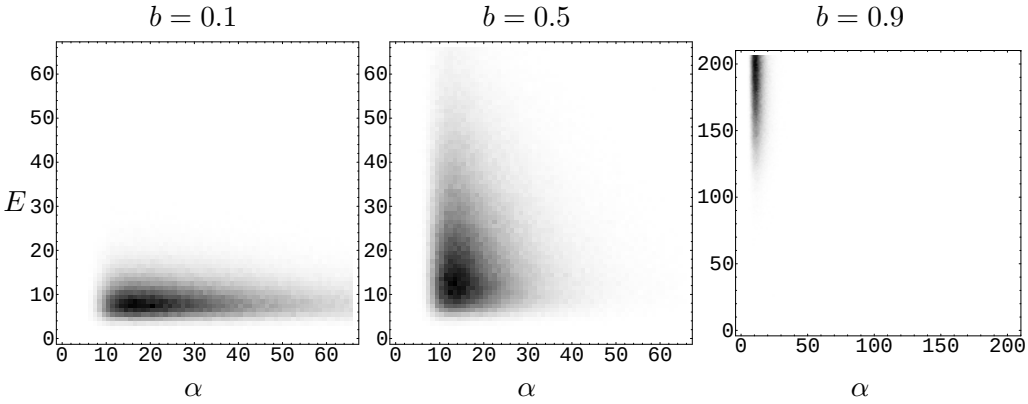

Figure 3.16: Place in (E, α) Space. The distance to the threshold is given in units of $\alpha_{\text{thr}}/100$ and $|E_{\text{thr}}|/100$

Mean Fitness

The mean fitness is a measure how good the protein is in resisting the destructive forces by mutation, or, to see it from the other side, how strong these forces are. In figure 3.15 the mean fitness in dependence of bias is shown. For ATPE, $\langle f \rangle$ is best for a b around 0.7 to 0.8 . There is a stabilizing effect by a small bias towards AT. For extreme b, the population is pushed to lower fitness (see figure 3.14).

(E, α) Space

For different b, the population lives at different places in the (E, α) space. A bias towards A and T pushes the population against the α-threshold and away from the E-threshold. A bias towards G and C accounts for lower E values, but α has to be kept up by selection. This effect is metered by w, as defined in 3.2.6. That the effect is stronger for b near

one that for b close to zero is expected. Putting hydrophobic amino acids in places with low PE component brings down α, but putting not hydrophobic amino acids in places with high PE component is punished in both E and α.

3.3.5 Summary of Influence of Bias

A genetic bias changes the probability for amino acids to occur. Due to the structure of genetic code, a bias to A/T increases the probability of very hydrophobic amino acids, while a bias to G and C accounts for a higher probability of less hydrophobic amino acids. This preference of either more hydrophobic or less hydrophobic amino acids influences the occupation of the places in the protein and its stability. Depending on the environment of the protein, a genetic bias can help it to remain stable, by helping it to keep either $|E|$ or α above threshold.

3.4 Response of a System to Change of Population Size

A change in population size changes the probability of fixation for a mutation. The larger the population size is, the more unlikely it is for deleterious and slightly deleterious mutations to get fixed. For neutral mutations, the acceptance ratio is only scaled

$$p_{\text{acc}} \propto \frac{1}{N} \tag{3.5}$$

A large population size prevents the system of going to states with lower fitness. As long as fitness stays up, results are independent of population size.

3.4.1 Changes on the DNA Level

The content of nucleotides is hardly dependent on population size, see figure 3.17. Even c_{2T}, which has the highest evolutionary pressure when a bias is present, does not change its behaviour.

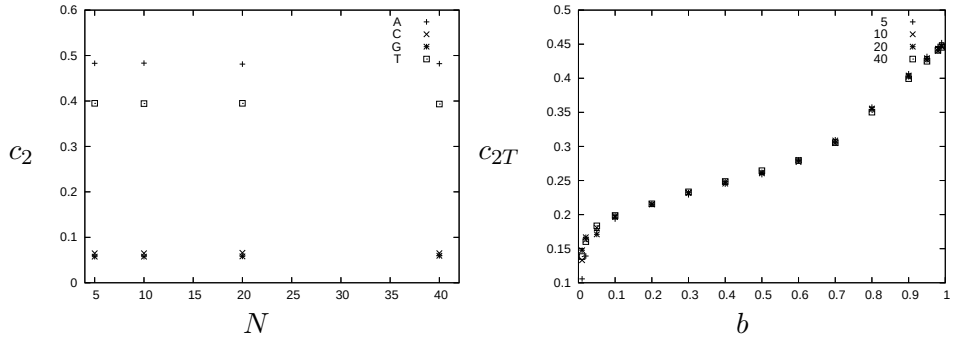

Figure 3.17: c_2 vs Population Size N for $b = 0.9$ (left), c_{T2} vs Bias b for Different Population Sizes (right)

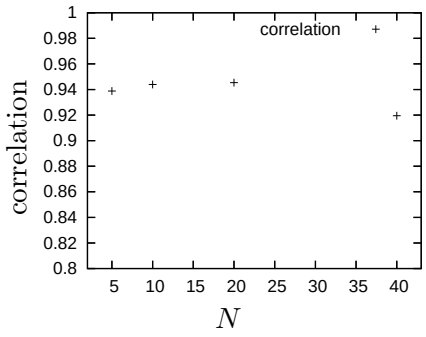

Figure 3.18: Correlation of β_i and $c_i/\langle c \rangle$ in Dependence of Population Size

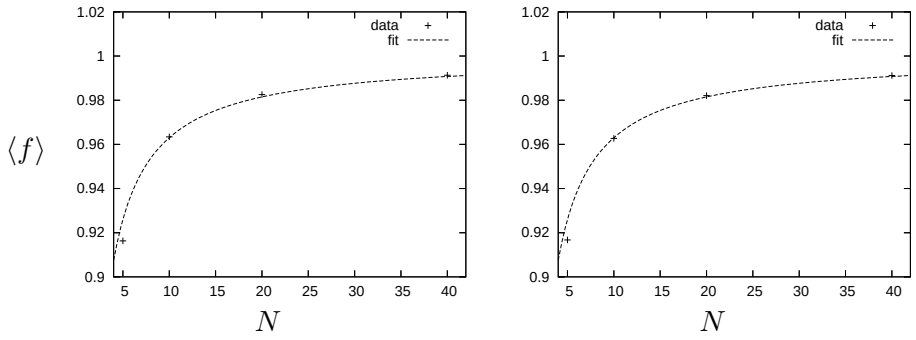

Figure 3.19: Mean Fitness for AT-Bias 0.1 and 0.9 vs Population Size

3.4.2 On the Amino Acid Level

The prediction of site-specific amino acid distributions is possible, independent of population size. The correlation of the exponential parameters β_i and the corresponding PE component are constant. Behaviour does not change significantly.

3.4.3 On the Population Level

Slightly deleterious mutations are repressed in a larger population. Therefore the mean fitness is expected to stay up for larger populations. Following Sella and Hirsh [17], the probability for the states of lower fitness, which bring down the mean fitness, to be populated is

$$p(s) \propto f(s)^N \tag{3.6}$$

Hence the mean fitness is expected to follow

$$\langle f \rangle^N = \text{const} \tag{3.7}$$

This behaviour is observed for every bias (see figure 3.19).

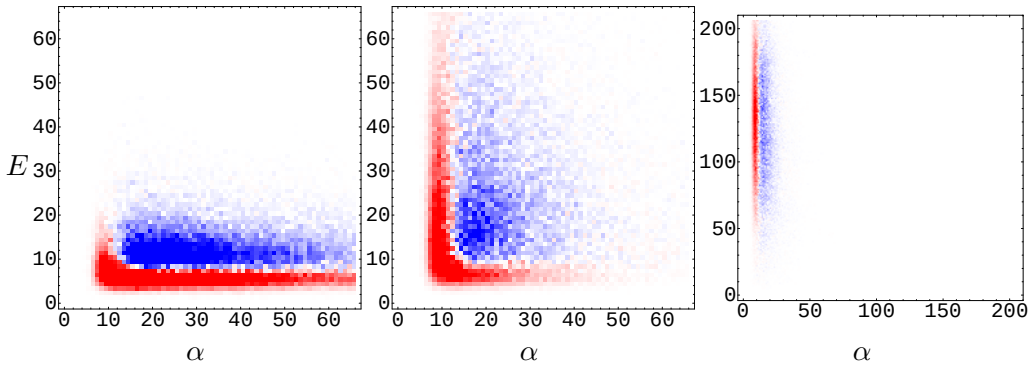

Figure 3.20: Difference between Population Size 10 and 20 for different AT-Bias

3.4.4 Location in the (E, α)-Space

The dependence of the location in (E, α) space on the bias, as described in 3.3.4, holds for all populations sizes. The distance to the threshold is different. The difference between the locations for different population sizes is illustrated in figure 3.20. In the blue area, the population is more often for $N = 20$. In the red area, the population is more often for $N = 10$.

3.4.5 Little Influence of Population Size

This system does a random walk in (E, α) space, which is disturbed if either $|E|$ or α drops below threshold. Effects of population size can only be seen when hitting the thresholds, because as long as the fitness does not change during the walk, the population size is just a scaling factor of acceptance ratio. Most of the time, the population has a fitness very close to one, so the random walk is disturbed only a few times. Even though the model of evolution is not completely neutral, the dependence on population size behaves like in neutral evolution. This is probably due to a too steep decay of the fitness function.

4 Discussion and Outlook

The simulation performed gives some insights into the interplay between protein thermodynamics and protein evolution. The evolution is simulated for the standard conditions, for some population from 5 to 200 and for different AT bias. Most of the time, the systems evolution is nearly neutral. To further improve knowledge about dynamics when hitting the thresholds in (E, α) space, further simulation are necessary. One possibility for this is to change the fitness-landscape. The decay of the fitness is very steep in the present simulations. A broader decay would prevent the proteins from just walking in the area where fitness is nearly one. Also some different steepness at the α- and the E-threshold will change the behaviour. Of great interest would be the question if the behaviour of the mean fitness, which showed that a small bias towards AT is of advantage, changes when changing the fitness landscape. This could give further insight in whether the optimal bias is a property of the system or a property of the fitness landscape.

In the results it is shown that the structure of the genetic code is one important determinant of the behaviour on the molecular (DNA) level. Therefore this code is an interesting aspect to focus on. Different codes could be simulated to answer the question, if the genetic code found in nature is optimized, and if so, in which respect.

A Appendix

A.1 Software Used

For the simulations, C++ code, planned in UML with Dia, compiled by g++ from the GNU Compiler Collection, organized by CVS, tested with the BOOST Test Framework and managed by Autoconf, Automake, and Libtool has been used. For random number generation, the 'gfsr4' generator from the GNU Scientific Library (GSL) was taken. For statistics and data analysis, Mathematica 5.2 was deployed. Visualization of data was done using Gnuplot. For typesetting LaTeX was used.

The websites of the Protein Databank (PDB) `http://www.rcsb.org/pdb` was used to download protein sequences and structures, as well as to visualize structures.

A.2 Amino Acids

Amino Acids abbreviations taken from [8], $h(a_i)$ taken from [10]:

name	3 let. abbr.	1 let. abbr.	$h(a_i)$
Alanine	Ala	A	0.1366
Arginine	Arg	R	0.0363
Asparagine	Asn	N	-0.0345
Aspartic acid	Asp	D	-0.1233
Cysteine	Cys	C	0.2745
Glutamic acid	Glu	E	-0.0484
Glutamine	Gln	Q	0.0325
Glycine	Gly	G	-0.0464
Histidine	His	H	0.0549
Isoleucine	Ile	I	0.4172
Leucine	Leu	L	0.4251
Lysine	Lys	K	-0.0101
Methionine	Met	M	0.1747
Phenylalanine	Phe	F	0.4076
Proline	Pro	P	0.0019
Serine	Ser	S	-0.0433
Threonine	Thr	T	0.0589
Tryptophan	Trp	W	0.2362
Tyrosine	Tyr	Y	0.3167
Valine	Val	V	0.4084

Bibliography

[1] M. Kimura and T. Otha, *Theoretical aspects of Population Genetics* (Princeton University Press, Princeton, New Jersey, 1971).

[2] U. Bastolla, A. Moya, E. Viguera, and R. C. van Ham, Journal of Molecular Biology **343**, 1451 (2004).

[3] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, in *Structural Approaches to Sequence evolution: Molecules, Networks, Populations*, edited by U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo (Springer, Berlin, 2006), Chap. The Structurally Constrained Neutral Model of Protein Evolution.

[4] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Journal of Molecular Evolution **56**, 243 (2002).

[5] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Physical Review Letters **89**, 208101 (2002).

[6] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Journal of Molecular Evolution **57**, 103 (2002).

[7] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, submitted to BMC Evolutionary Biology (unpublished).

[8] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution, Second Edition* (Sinauer Associates, Inc., Massachusetts, USA, 2000).

[9] U. Bastolla, J. Farwerand, E. W. Knapp, and M. Vendruscolo, Proteins **44**, 79 (2001).

[10] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Proteins: Structure, Function, and Bioinformatics **58**, 22 (2005).

[11] U. Bastolla, H. Roman, and M. Vendruscolo, European Physical Journal B **15**, 385 (1999).

[12] B. Derrida, Physical Review B **25 5**, 2613 (1981).

[13] U. Bastolla and L. Demetrius, Protein Engineering, Design & Selection **18**, 405 (2005).

[14] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Gene **347**, 219 (2005).

[15] M. Porto, H. E. Roman, M. Vendruscolo, and U. Bastolla, Molecular Biology and Evolution **22**, 630 (2005).

[16] M. Porto, H. E. Roman, M. Vendruscolo, and U. Bastolla, Molecular Biology and Evolution **22**, 1156 (2005).

[17] G. Sella and A. E. Hirsh, Proceedings of the National Academy of Sciences **102**, 9541 (2005).

Acknowledgments

I would like to thank Prof. Markus Porto for an at all time excellent support as well as for creating a very pleasant working environment. Furthermore, I would like to thank Dr. Ugo Bastolla and Prof. Markus Porto for introducing me into the topic and for lots of mails and extensive discussions about the subject of this thesis, as well as about related questions.

I would like to thank Kathrin Weyer for lots of support in the last weeks of work on the thesis.

Furthermore, I would like to thank Dennis Ratzke for discussions which were very clarifying concerning some points, and Johannes Größchen who proofread this document very carefully.

Erklärung zur Bachelor Thesis gemäß §23 Abs. 7 APB:

Hiermit versichere ich, die vorliegende Bachelor Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus den Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, 15. März 2006

Impressum	
Author:	Andreas Buhr, Karlstrasse 18, D-63225 Langen
Satz:	LaTeX
E-Mail	andreas at andreasbuhr.de
Web	www.andreasbuhr.de