Inverse folding of Proteins with Knowledge Based Potentials of Mean Force

Diplomarbeit

zur Erlangung des
akademischen Grades

Magister rerum naturalium

an der Formal und Naturwissenschaftlichen
Fakultät der Universität Wien

eingereicht von

Aderonke BABAJIDE

Institut für Theoretische Chemie

Wien, im April 1996

Abstract

Inverse folding of proteins with *potentials of mean force* is one of the most promising current approaches to the protein folding problem. These potentials are derived from the data contained in known protein structures in the large Protein Data Banks (e.g. Brookhaven Protein Data Bank). These potentials are based on two basic *assumptions*:

(1) Proteins fold into a thermodynamic ground state.

(2) The frequency $\phi(I)$ of a certain interaction I in the data base of all (known) protein structures is related to the energy contribution $E(I)$ of I by means of Boltzman's law.

In a series of papers Sippl and coworkers (Center for Applied Molecular Engineering, University of Salzburg) [27, 28, 29] showed that a good potential of mean force can be derived from the frequencies of amino acid residues in structure data bases. Exploiting this data, they developed the `PROSA` program package which allows us to calculate a cost function (*z-score*) for a given amino acid sequence on a certain protein structure. In this work, the existence and the extension of neutral paths and networks on a variety of protein structures were studied using the `PROSA` program. Furthermore, this inverse folding approach allowed us to study the number of amino acids necessary to model native-like protein structures, i.e., we studied the feasibility of structure formation with restricted alphabets. We found that it is possible to generate random amino acid sequences with native like *z-scores* and sequences with even better *z-scores* by *adaptive walks*. This was even possible with restricted sets of amino acids, given the right combination of hydrophilic and hydrophobic amino acids. Neutral paths extended to the length of the amino acid sequence at *z-scores* equal to the wild-type and up to six standard deviations better. Preliminary studies of hydrophilic−hydrophobic patterns in the generated sequences have not yielded any discernible patterns so far. There seems to be no bias in the substitution frequency of the amino acids during the adaptive walks and during the search for neutral neighbors.

Zusammenfassung

Die inverse Faltung von Proteinen unter Verwendung von Potentialen der mittleren Kraft ist einer der derzeit vielversprechendsten Ansätze zur Lösung des Proteinfaltunsproblems. Diese Potentiale werden aus Proteinstrukturdaten entwickelt, welche man in den großen Protein-Datenbänken (z.B. Brookhaven Protein Data Bank) findet. Sie basieren auf den folgenden grundlegenden *Annahmen*:

(1) Gefaltete Proteine befinden sich in einem thermodynamischen Grundzustand.

(2) Die Frequenz $\phi(I)$ einer bestimmten Interaktion I im Datenpool aller bekannten Proteinstrukturen ist mit der Energieverteilung $E(I)$ von I über das Boltzmann-Gesetz verknüpft.

In einer Reihe von Publikationen zeigten Sippl, *et al* (Center for Applied Molecular Engineering, University of Salzburg) [27, 28, 29], daß ein gutes Potential der mittleren Kraft aus der Frequenz von Aminosäuren in Strukturdatenbanken gewonnen werden kann. Unter Verwendung der dort enthaltenen Daten entwickelten Sie das `PROSA` Programmpaket. Dieses Programm ermöglicht uns die Berechnung einer Kostenfunktion (*z-score*) für eine bestimmte Aminosäuresequenz auf einer vorgegebenen Proteinstruktur. In dieser Arbeit wurde die Existenz und die Ausdehnung von neutralen Pfaden und neutralen Netzen auf einer Anzahl von Proteinstrukturen untersucht. Hierfür wurde das `PROSA` Programm verwendet. Desweiteren versuchten wir herauszufinden wieviele Aminosäuren notwendig sind um Aminosäuresequenzen zu
entwickeln deren *z-score* dem nativen entspricht, d.h. wir untersuchten die Verwendbarkeit von limitierten Alphabeten. Es stellte sich heraus, daß es mit Hilfe von *Adaptive Walks* möglich ist Aminosäuresequenzen zu modelieren deren *z-score* dem der nativen Sequenz entspricht, bzw. solche mit wesentlich besseren *z-scores.* Mit der richtigen Kombination von hydrophilen und hydrophoben Aminosäuren war dies auch bei limitierten Alphabeten möglich. Bei *z-scores*, die den nativen ähnlich waren, bzw. bis zu sechs Standardabweichungen besser, entsprach die Länge der neutralen Pfade der Länge der Aminosäuresequenzen. Vorläufige Untersuchungen der hydrophil−hydrophob-Muster ergaben bis jetzt keine erkennbaren Muster. Die Frequenz der Substitution der Aminosäuren bei den Adaptive Walks bzw. bei der Suche nach neutralen Nachbarn scheint keinem Bias zu unterliegen.

1. Introduction

The protein folding problem is one of the most interesting problems of contemporary biology. The solution to this problem would pave the way for a great number of scientific and technological applications. During the past decades the folding problem resisted the attacks of intense theoretical research. However, during the last few years the invention of new strategies has brought us a few steps closer to a solution. These strategies are based on the analysis of known three-dimensional structures using methods borrowed from statistical physics.

One of the most promising approaches seems to be the use of knowledge-based potentials of mean force, derived from the data contained in known protein structures in the large Protein data banks (e.g. Brookhaven protein data bank; (`URL: http://pdb.pdb.bnl.gov/`). Sippl and co-workers [17, 27] argue that a good potential of mean force can be derived from the frequencies of amino acid residues in structure data bases. These potentials are based on two basic *assumptions*:

(1) Proteins fold into a thermodynamic ground state, i.e., the conformation of a proteins minimizes a potential function.

(2) The frequency $\phi(I)$ of a certain interaction I in the data base of all (known) protein structures is related to the energy contribution $E(I)$ of I by means of Boltzman's law.

One should bear in mind that both assumptions are by no means obvious *a priori* (see section 4.1). Although the prediction of protein conformation from amino acid sequences has remained an unsolved problem in spite of all efforts, the inverse folding problem of finding a sequence that will adopt a predefined structure as its native conformation may be more tractable for several reasons:

At a structural resolution comparable to that of a ribbon diagram of a protein many sequences will adopt the same structure, and optimization of sequences in high dimensional sequence spaces should be easy compared to optimization of structures in 3D. Reliable recognition of correctly folded proteins should be easier than structure prediction and is all that is needed. Solving the inverse folding problem is not only a prerequisite for rational design of functional proteins, but also allows to study sequence-structure relations of proteins. Knowledge-based potentials of mean force have been used successfully to identify a proteins native fold among a large set of possible conformations [17, 28, 29]. Since the score assigned to a given conformation correlates well with its distance to the native fold (see Figure 10), one might, conversely, interpret the difference in score between a wild type and some test sequence as a measure of distance between their respective native folds. The scores obtained from the mean force potentials would then allow us to search protein sequence space in a similar way as has been done for RNA secondary structures [9, 26], without having to tackle the problem of protein structure prediction. The only restriction is that only proteins with known structure can serve as reference.

Sippl and co-workers have shown in a series of papers [4, 17, 27, 28, 29] that the potential of mean force W, or rather the rescaled quantity

$$z(p,Q) := \frac{W(p,Q) - \overline{W}_p}{\sigma_p},$$

where $\overline{W}_p$ and σ_p are the mean and standard deviation of $W(p,Q)$ when Q runs over all conformations in a database of known protein structures, can be used to identify the native fold P among a large set of possible conformations Q.

Conversely, this *z-score* can be used as an approach to *inverse folding*: Given a fixed backbone conformation Q, one could search for sequences p that give z-scores $z(p,Q)$ at least as low as the z-score of the native sequence q. Of course, only structures that are already in the database can be searched for.

In the case of RNA molecules, and if one is willing to accept secondary structures, i.e., base pairing patterns, as a suitable (coarse grained) description of the structures, one can actually compute the structure of minimum free energy for (in principle) arbitrary sequences [18, 23, 30, 31]. These algorithms are based on a simple thermodynamic model of RNA (secondary) structures, for which the majority of parameters have been measured directly on small oligonucleotides [10]. The simplicity of the energy model and the relatively small number of contributions in a given sequence allow this approach to be applied successfully.

Using a brute force computational approach a number of unexpected results on the global properties of the sequence-structure map of RNA's have been obtained [7, 8, 9, 26]. The highlights of these studies are:

(i) There are many more sequences than structures, hence many sequences fold into the same structure.

(ii) The distribution of the number of sequences folding into the same structure follows a Zipf-like distribution, i.e., there are few very common structures and many different very rare structures.

(iii) The sequences folding into a common structure are distributed randomly throughout sequence space. No clustering is visible.

(iv) The sequences folding into a common structure form extremely extended *neutral networks*, i.e., there are pathways consisting of sequences that fold into same structure which extend through all sequence space.

(v) The distance from a random sequence to a sequence that folds into a desired structure is short compared to the maximum distance in sequence space.

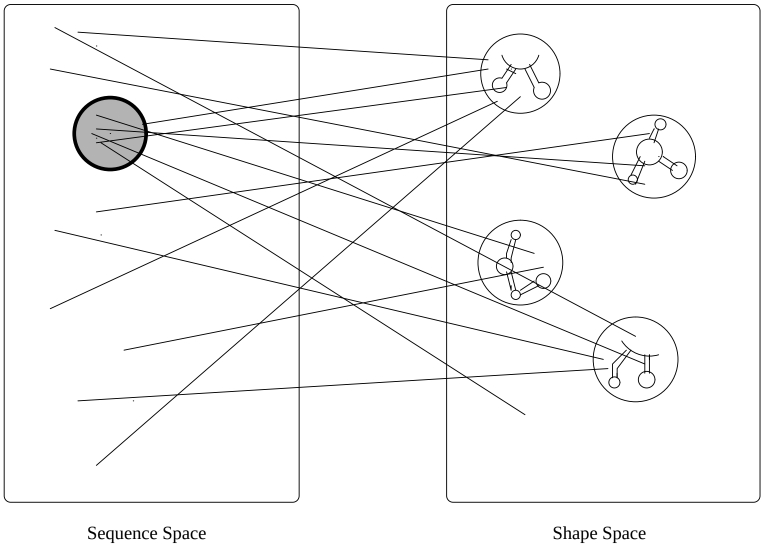

Figure 1: Sequence − Structure relation in the case of RNA secondary structures: Sequences folding into a particular structure can be found anywhere in sequence space. Such sequences can be connected by extended nets of structurally neutral neighbors.

The development of knowledge-based potentials of mean force enables us to ask similar questions for proteins:

(i) How are the sequences folding into common structures distributed throughout sequence space?

(ii) Is it possible to create random sequences of amino acids, that fold into a given structure yielding the same or better z-scores than the wild-type sequence?

(iii) Do sequences which fold into common structures create neutral paths or networks throughout sequence space?

(iv) Is it possible to create native-like protein structures using a restricted set of amino acids such as ADLG (Alanine, Aspartate, Leucine and Glycine)?

(v) How many amino acids are needed to build typical structures, i.e. is it possible to create native-like protein structures using only an hydrophobic and an hydrophilic amino acid?

Recently a number of groups (for details see [5, 20, 25]) have begun to approach some of these questions experimentally. In this work we tried to find answers with the help of computer experiments using the `PROSA` program developed by Sippl, *et al.* [29], (see section 4.3).

2. Protein Structure

Proteins play important roles in virtually all biological processes. The wide range of their activity and their significance are exemplified by the following functions:

1. Enzymatic catalysis
2. Transport and storage
3. Coordinated motion
4. Mechanical support
5. Immune protection
6. Generation and transmission of nerve impulses
7. Control of growth and differentiation

Proteins consist of one or more polypeptide chains built from a repertoire of twenty α-amino acids. An α-amino acid consists of an amino group, a hydrogen atom, and a distinctive R group (side chain) all bonded to a carbon atom (the α-carbon) which is adjacent to the carboxyl group.

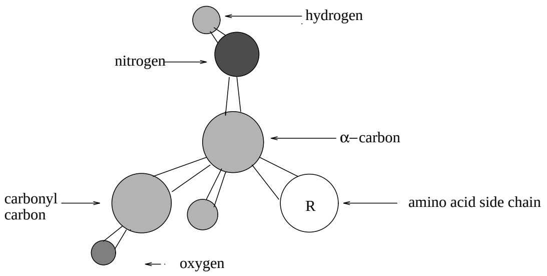

Figure 2: An α-amino acid

Amino acids are amphiphatic molecules, in solution at a neutral pH they are dipolar ions rather than unionized molecules.The tetrahedral array of four different groups about the α-carbon atom accounts for the optical activity of amino acids. Of the two possible isomers only L-amino acids are constituents of proteins.

The side chains of the twenty different amino acids vary in size, shape, charge, hydrogen bonding capacity and chemical reactivity. The sequence of amino acids in the polypeptide chain(s) is also called the primary sequence of the protein and specifies the three dimensional structure of the protein.

Table 1. The 20 Standard Amino Acid Residues

Residues	Symbols	H	P	+	−	C	s	ar	al
Alanine	Ala A	x					x		
Arginine	Arg R		x	x		x			
Asparagine	Asn N		x				x		
Aspartate	Asp D		x		x	x	x		
Cysteine	Cys C	x					x		
Glutamate	Glu E		x		x	x			
Glutamine	Gln Q		x						
Glycine	Gly G	x					x		
Histidine	His H	x	x	x		x		x	
Isoleucine	Ile I	x							x
Leucine	Leu L	x							x
Lysine	Lys K	x	x	x		x			
Methionine	Met M	x							
Phenylalanine	Phe F	x						x	
Proline	Pro P						x		
Serine	Ser S		x				x		
Threonine	Thr T	x	x				x		
Tryptophane	Trp W	x	x					x	
Tyrosine	Tyr Y	x	x					x	
Valine	Val V	x					x		x

Classifications: H ... hydrophilic, P ... polar, + ... positive, − ... negative, C ... charged, s ... small, ar ... aromatic, al ... aliphatic [32].

The amino acids in the polypeptide chain are linked by peptide bonds i.e. the α-carboxyl group of one amino acid is joined to the α-amino group of another amino acid. The carbon-nitrogen bond has partial double bond character, therefore the peptide group is a rigid planar unit. As a consequence, rotation can only take place about the bonds on either side of the rigid peptide unit.

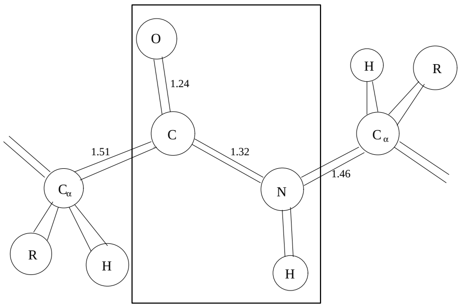

Figure 3: The peptide bond. Standard bond distances are given in Å.

Rotation about the two single bonds is described by the dihedral angles ψ and ϕ. ψ refers to rotations about the C_α — C single bond while ϕ refers to rotations about the C_α — N single bond. In a fully stretched out polypeptide chain $\phi = \psi = 180°$. The conformation of the main chain of a protein is completely defined when ψ and ϕ are specified for each residue in the polypeptide chain.

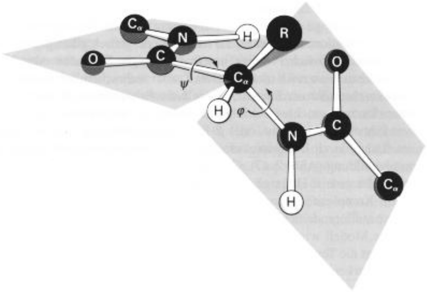

Figure 4: Definition of ϕ, ψ.

The polypeptide chain folds into regular structures stabilized by H-bonds. Three structural features are prominent in protein secondary structure: Helices, β Pleated Sheets and Turns.

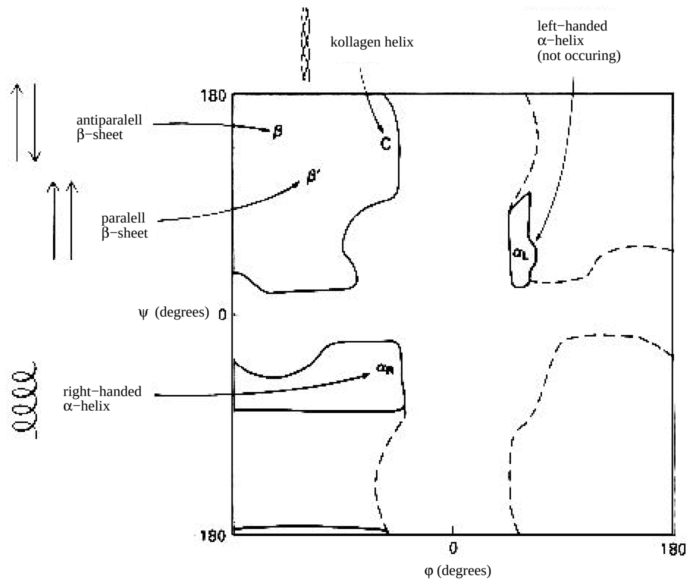

Figure 5: Ramachandran plot; poly-L-*ala*

Secondary structures are of particular interest for the understanding of the mechanism of protein folding. They are also important for the theory and prediction of protein structures. In the following we will discuss the main structural elements of proteins.

- Helices

The prediction of α-Helices as essential structural elements in proteins made by Linus Pauling and Robert Corey turned out to be a milestone in the understanding of biopolymeres. It is a right-handed helix (figure 6) of the polypeptide-chain. Each amino acid residue in an α-helix forms a H-bond between its carbonyl-group and the amino-group of the fourth next amino acid:

$$k \to k+4; \qquad k = 1, 2, 3, \dots .$$

By creating an α-helix, the polypeptide chain transforms into a more compact and more stable form.

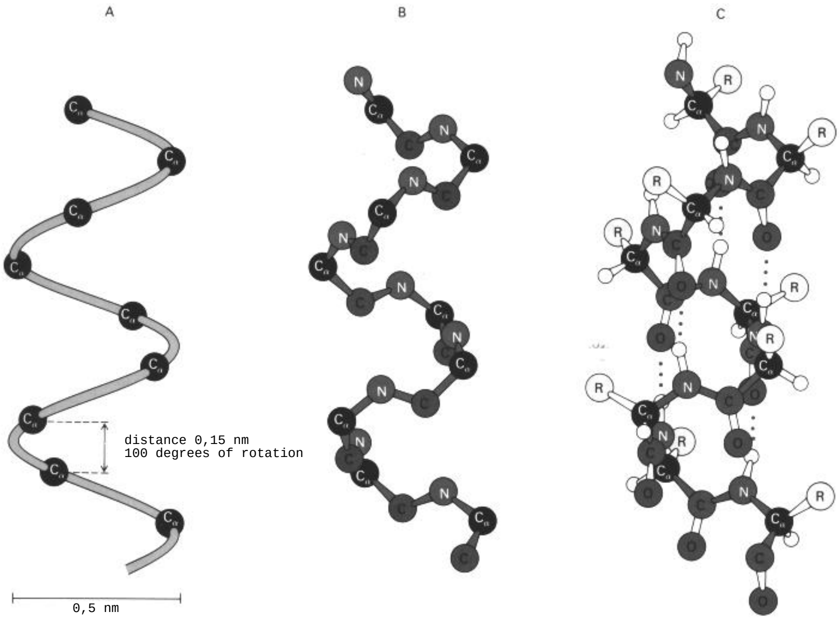

Figure 6: α-Helix: left: only α-C-atoms(C_α). middle: C_α+ N + C of the backbone. right: total helix

Helical structures can be described in several ways. One of them utilizes the above mentioned dihedral angles ϕ and ψ and the angle ω which is approximately 180° in most structures.

Another characterization is based on the number of amino acids n necessary to complete a full turn of the helix. Furthermore, m which counts the number of atoms in the ring created by an CO — HN bond, has to be given. The helix is fully described by n_m. The smallest possible ring can be found in the 2.2_7-helix. In this case there are no "X".

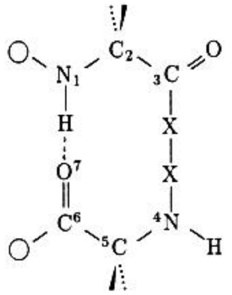

Figure 7: H-bond in an α-Helix. The continuation of the polypeptide-chain is indicated by the two circles. "X" symbolizes amino acids in between the H-bond.

There is another important parameter: h. It describes the contraction of the polypeptide chain in translational direction. It is usually measured in Ångstroms (Å).
Questions concerning protein stability can be answered by (ϕ, ψ)-potential-fields, which are very similar to Ramachandran-plots.
In addition to right handed helices there are left handed helices:

$$\psi_{lh} = -\psi_{rh} \qquad \text{and} \qquad \phi_{lh} = -\phi_{rh}$$

These helices are *no* mirror images of their right handed counterparts, because all amino acids, except *gly*, are *chiral* and only the L-isomers are represented.

- β Pleated Sheets

In contrast to the α helix, the β pleated sheet is a *non* local structural unit. The polypeptide chain in this structural element is almost fully extended. It is stabilized by H-bonds between NH and CO groups in *different* polypeptide chains. Adjacent chains in a β pleated sheet can run in the same direction *(parallel β sheet)* or in opposite directions *(antiparallel β sheet)*.
In general these structures are called β-structures The molecular properties are shown in figure 8. The side chains are alternately orientated to both sides. There is only one exception: L-*pro*, because there is no hydrogen bond to the nitrogen and in addition it cannot rotate to the required angles (ψ, ϕ).

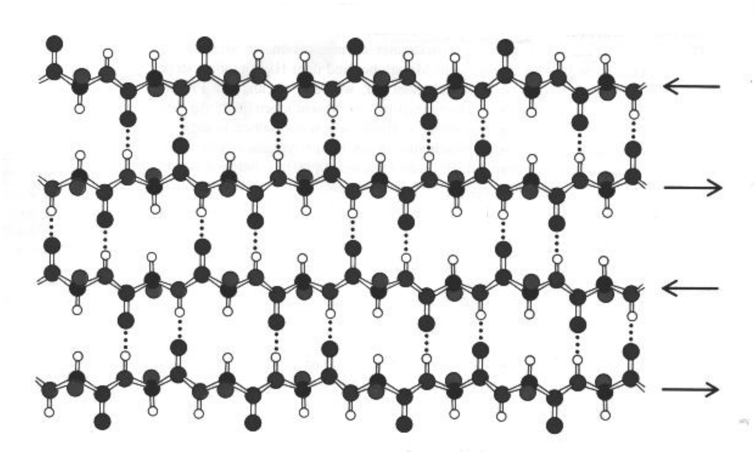

Figure 8: Anti-parallel β-sheet

- Turns

Globular proteins are of approximately spherical shape, the α-helixes and β-sheets they contain cannot be longer than their diameter. Consequently the polypeptide chain has to change its direction with the help of *hairpin- or β- turns.* ‘Sharp” turns often occur in β-sheets. They consist out off four amino acid-residues. There is a H-bond between the CO of the first and NH of the fourth next amino acid. The following restrictions concerning the occurance of certain amino acid residues at certain positions can be observed with the different types of turns:

I all residues are allowed on positions 1-4 except *pro* in position 3

I’ positions 2 and 3 must be *gly*

II position 3 must be *gly*,

II’ position 2 must be *gly*

III is part of a 3.0_{10}-helix (no further restrictions)

III’ positions 2 and 3 must be *gly*

IV must have *pro* in position 3 and a cis-peptide bond

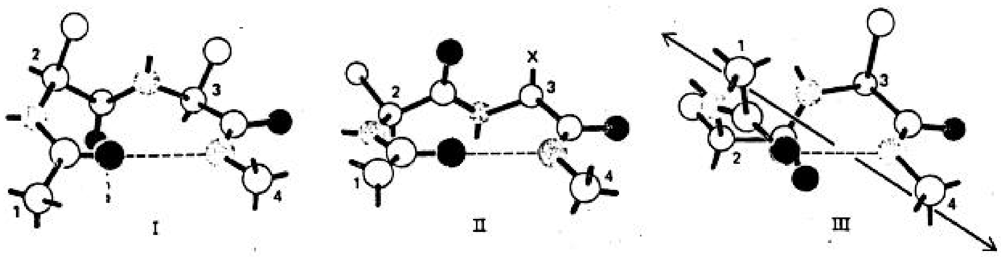

Figure 9: Most frequent hairpins. X marks the position where an H-atom is required. Type III shows a strong similarity to a 3.0_{10}-helix (axes marked)

- Ω-turns

These are turns with a length of approximately 6-16 residues, they combine other secondary structure elements and have an end-to-end distances of 3.7-10 Å. Furthermore, the residue-distribution differs from the α-helix and the β-sheet. In particular *gly* and *pro* are very frequent.

3. Globular Proteins

Most biological proteins have characteristic native folds. In physiological conditions the native structure forms spontaneously. The protein's folded structure is a function of its amino acid sequence and its natural environment. In the case of soluble globular proteins the in vivo environment is generally an aqueous solution of various ingredients. The amino acid sequence defines the molecular identity of a protein. The study of the biological role, molecular mechanism, catalysis, molecular interactions, binding of effector molecules, and many more important features of individual proteins require a knowledge of their three dimensional structures. In this work we study four globular protein structures in detail.

3.1. Thioredoxin

Thioredoxin is an electron carrier protein. It acts as an electron donor in the reduction of ribonucleotides and plays an important role in controlling the dark reaction of photo-synthesis. It controls the activities of various enzymes in many kinds of cells by reducing disulfide bonds. The active form of thioredoxin contains two cystein which are oxidized to form a disulfide bond when thioredoxin activates other enzymes. Thioredoxin is reactivated by reduction of the disulfide bond by ferredoxin. The Thioredoxin used for the following studies is that of Escherichia coli. It contains 108 amino acids.

The secondary structure (as shown in Figure 10) contains five α-Helixes (H1 Residue 11 to 17, H2: 32 to 49, H3: 59 to 63, H4: 66 to 70, H5: 96 to 107), five β-sheets (Residues 3 to 8, 29 to 32, 53 to 59, 76 to 82, 86 to 92) and 12 Turns (see Figure 9). The active site is located at the amino-terminus of the second alpha-helix. It contains a disulfide bridge between Cys32 and Cys35. The structure used for the following calculations is `2trxA.pdb` (Brookhaven Protein Data Bank), resolved at a resolution of 1.68 Å.

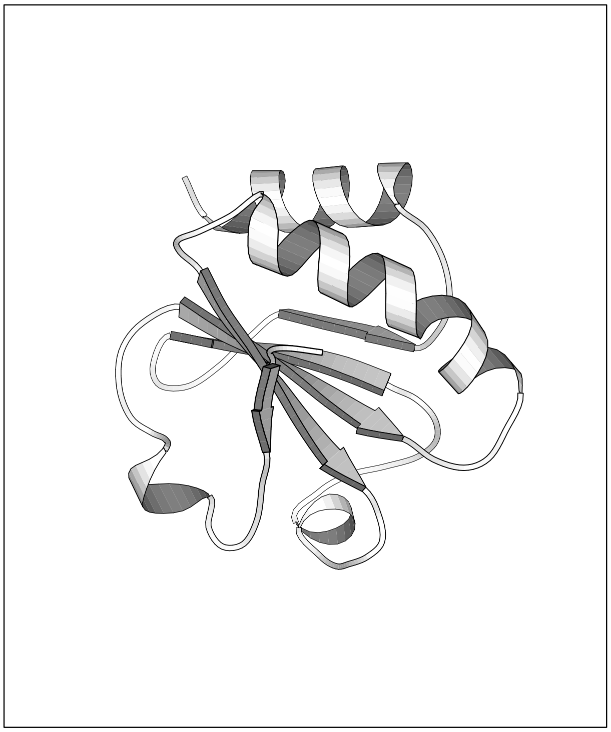

Figure 10: Thioredoxin from Escherichia Coli. Wild type Sequence:
SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGI
PTLLLFKNGEVAATKVGALSKGQLKEFLDANLA

3.2. Crambin

Crambin is a plant seed protein from abyssinian cabbage (Crambe abyssinica). It is the seed-specific thionin. Contrary to most thionins of higher plants which are toxic to various bacteria, fungi, and animal and plant cells, it exhibits no toxicity. Crambin has no net charge, it is very hydrophobic. It contains 46 amino acids.

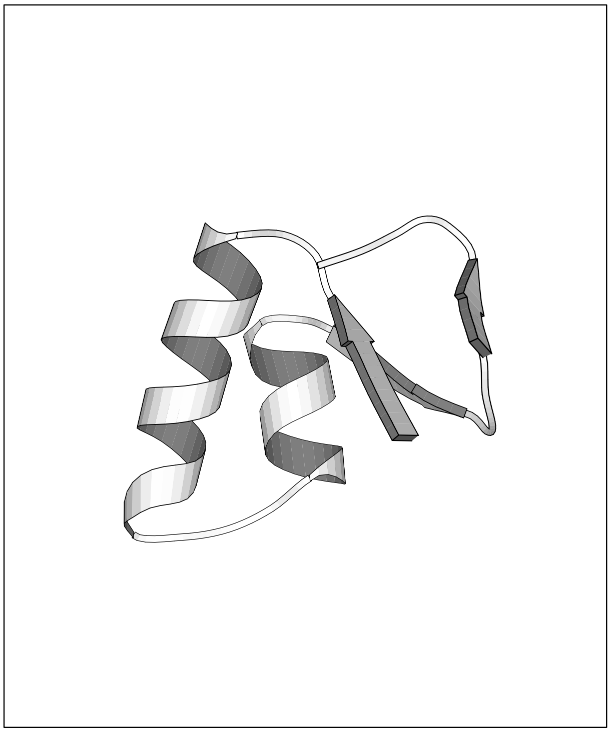

Figure 11: Crambin from Abyssinian Cabbage Seed. Wild-type sequence:
`TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN`

The structure used for the following calculations is `1cbn.pdb` (Brookhaven Data Bank), resolved at a resolution 0.83 Å. It contains two α-Helixes (H1 Residue 7 to 19, H2: 23 to 30), 3 β-sheets (Residues 1 to 4, 32 to 35, 39 to 41) and 2 Turns. There are three disulfide bonds: between Cys 3 and 40, 4 and 32, 16 and 26 (see Figure 11).

3.3. Ubiquitin

Ubiquitin is a small protein present in all eucaryotic cells. It plays an important role in tagging proteins for destruction. This protein is highly conserved in evolution: yeast and human ubiquitin differ at only 3 of 76 residues. The carboxyl-terminal glycine becomes covalently attached to the ϵ-amino group of lysine residues of proteins destined to be degraded. Ubiquitin from human erythrocytes was used for the following calculations.

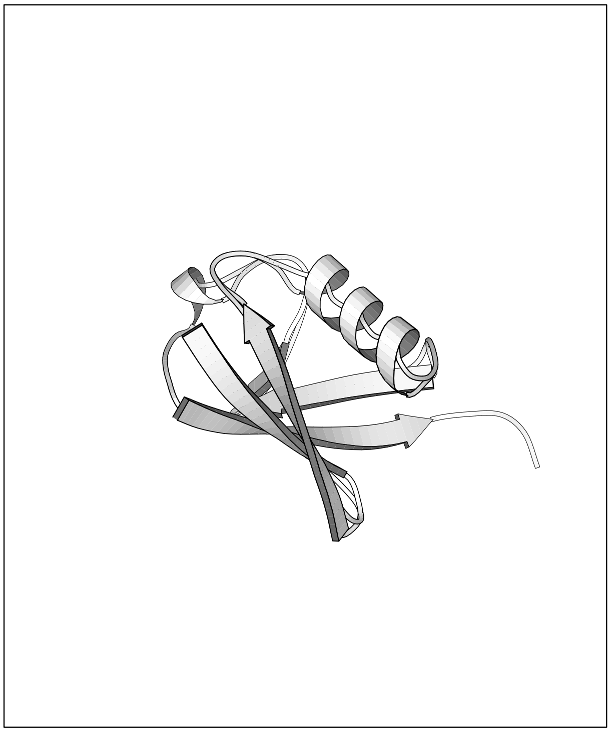

Figure 12: Ubquitin from Human Erythrocytes. Wild-type sequence:
`MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDY`
`NIQKESTLHLVLRLRGG`.

The structure used for the following studies was `1ubq.pdb` (Brookhaven Protein Data Bank), resolved at 1.8 Å. It contains two α-Helixes (H1 Residue 23 to 34, H2: 56 yo 59), five β-sheets (Residues 1 to 7, 10 to 17, 40 to 45, 48 to 50, 64 to 72) and 9 Turns (see Figure 12).

3.4. Lysozyme

Lysozyme is an enzyme capable of dissolving certain bacteria (lysis) by cleaving the polysaccharide component of their cell wall. It is a relatively small enzyme. The lysozyme from chicken egg white which was used for the following calculations, is a single ploypeptide chain of 129 residues. This highly stable protein is cross-linked by four disulfide bridges: between Cys 6 and 127, 30 and 115, 64 and 80, 76 and 94. The active site contains Asp 52 and Glu 35.

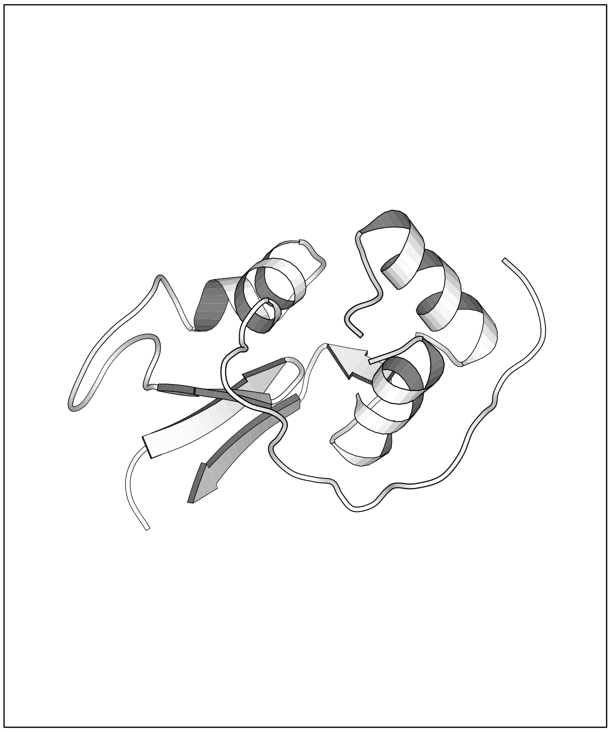

Figure 13: Lysozyme from Hen Egg White. Wild-type sequence:
`KKLGRCELAAAMKRHGLQNERGLSMGNWVCAAAFESNFNTQATNRNTDGSTDYTFLQINSRWW`
`CNDGRAPGSRNLCGIPCSALLSSDITASVNCAVKIYSDGNGCNIMVAWRNRCKGTDEQRWIRGCRL`

The structure used for the following studies was `1lyz.pdb` (Brookhaven Protein Data Bank), resolved at 2.0 Å. It contains four α-Helixes (H1 Residue 5 to 15, H2: 25 to 35, H3: 80 to 84 and H4: 89 to 96, five β-sheets (Residues 1 to 3, 38 to 40, 42 to 46, 50 to 54 and 57 to 60) and 11 Turns (see Figure 13).

4. Protein Inverse Folding

4.1. Knowledge-Based Potentials of Mean Force

The problem of biopolymer folding can be paraphrased in terms of the *combinatory map* from the sequence space $\mathcal{Q}^n_\kappa$ consisting of all sequence of length of n built from the alphabet of κ different types of monomers to the shape space $\mathcal{S}$

$$S: \quad \mathcal{Q}^n_\kappa \to \mathcal{S}, \quad p \mapsto S(p).$$

In principle this function is defined for all sequences. In the case of RNA molecules, and if one is willing to accept secondary structures, i.e., base pairing patterns, as a suitable (coarse grained) description of the structures, one can actually compute the structure of minimum free energy for (in principle) arbitrary sequences [31, 30, 23, 18]. These algorithms are based on a simple thermodynamic model of RNA (secondary) structures, for which the majority of parameters has been measured directly on small oligonucleotides [10]. The simplicity of the energy model and the relatively small number of contributions in a given sequence allow this approach to be applied successfully. In the case of proteins the situation is much less fortunate. The mapping S is not computable within the framework of present-day algorithms and/or computer technology. However, the solution of the inverse folding problem alone would be sufficient to study the questions raised in the introduction.
An approach analogous to RNA based on minimizing an energy function faces insurmountable problems in the case of protein folding: (1) while there is essentially one predominant type of interactions in nucleic acids, namely base pair stacking (which is highly specific and involves only a small number of monomers), the dominating energy contributions in protein folding originate from more or less unspecific hydrophobic interactions, which may involve a large number of monomers. Hence there is a large number of relatively small contributions to the energy of folding, implying that the individual energy contributions have to be known even more accurately than in the RNA case [6]. In reality the hydrophobic contributions are hardly known or measurable independently of the protein in question at all. Consequently we cannot reasonably use the potential function $W(p, Q)$ for computing

the sequence-structure map S. All we know *directly* about S for proteins is the small list of sequences for which the structures are in a data base.
In the following we argue that this knowledge together with a special kind of knowledge-based potential function for proteins can in fact be used to obtain results on the global features of S despite the fact that we cannot solve the protein folding problem.
For simplicity of the discussion we will restrict ourselves to the C_α-backbone of the proteins. Let $\mathbf{x}_i$ denote the spatial coordinates of the C_α atom number i along the chain. For a sequence p we will use the notation p_i to denote the monomer at position i. The Euclidean distance between two C_α atoms is given by:

$$d_{ij} := \|\mathbf{x}_i - \mathbf{x}_j\|$$

Sippl and co-workers [17, 27] argue that a good potential of mean force can be derived from the frequencies of amino acid residues in structure data bases. As mentioned earlier, these potentials are based on two basic *assumptions*:

(1) Proteins fold into a thermodynamic ground state, i.e., the conformation of a proteins minimizes a potential function.

(2) The frequency $\phi(I)$ of a certain interaction I in the data base of all (known) protein structures is related to the energy contribution $E(I)$ of I by means of Boltzman's law:

$$\phi(I) = \frac{1}{Z} \exp\left(-E(I)/RT\right),$$

where Z is the partition functions defined as

$$Z = \sum_J \exp\left(E(J)/RT\right).$$

One should bear in mind that both assumption are by no means obvious *a priori*. Claim (1) leads to Levinthal's [21] paradox, as in general finding the global optimum of a complicated potential function will require much more time than folding of a protein requires in nature. Claim (2) implicitly assumes that (a) the energy contributions built into the potential function (pair energies and surface terms in the present case) are in fact the dominant contributions, and (b) that the frequency of the types of interactions averaged over a large sample of proteins is untouched by evolutionary selection.

Nevertheless these assumptions are reasonable: It is unlikely that protein structures even if determined by kinetic folding pathways could be *far* away from the ground state; hence protein structures will be very low lying states if not the global minima of the potential surface, and hence can sensibly be used to estimate the potentials. Claim (2) is essentially a "maximum entropy" assumption, and we simply do not have any evidence at present that it is violated.

The final result is a *potential of mean force* for all pairs (p, Q) consisting of a sequence p and a backbone conformation Q (as described by the set of Euclidean coordinates of Q). In the present case it is of the form

$$W(p,Q) = \sum_{i<j} W[p_i, p_j, |i-j|; d_{ij}] + \sum_i W_s[p_i; \chi(i)].$$

The additive pair-contributions $W[a, b, k; r]$ depend on the pair of amino acid residues (a, b), their separation k along the chain, and their Euclidean distance r in the conformation Q. Bowie, Eisenberg and co-workers [2, 3, 22] have demonstrated that the solvent exposure of an amino acid can be used to model the energetic features of solvent-protein interactions. Consequently, the potential of mean force $W(p, Q)$ contains surface terms $W[a, \chi]$ depending on the amino acid residue a and the number χ of protein atoms within a sphere of radius R_0 centered at the C^α atom of a. The parameter χ serves as a (crude) quantitative measure for the surface-exposure of the residue a. It is worth noticing at this point that the energy contributions are defined in terms of two classes of parameters: the first set depends only on the sequence (amino acid residues and separation along the sequence), while the other depends only on the spatial conformation (Euclidean distance and surface exposure).

This approach to knowledge-based potentials is by no means the only one, other groups have developed other knowledge-based, empirical potentials of mean force which are not based on Boltzman's law [1, 3, 12, 13, 14]. These approaches fall into different groups. The first group considers the observed frequency with which the distance between pairs of amino acids appear within one or more distance bins, in known crystal structures. This approach is limited to considering pair interactions [12], The second group constructs a definition of an "environment" for an amino acid based on the properties of the amino acids (polarity, secondary structures, etc.). These characteristics are coarsely binned so one can approximate by frequency counting the conditional probability of a single amino acid appearing in an environment. Alan Lapedes *et al.* [14] have developed a formalism to construct "contact potentials". This formalism allows the introduction of machine learning techniques, such as *Neural Networks* which can efficiently include higher order interactions without the explosion of parameters. They employ hidden neurons to detect correlations higher than second order and they do not rely on frequency counting to approximate probability distributions.

4.2. The z-Score as a Structure Distance

Sippl and co-workers have shown [4, 17, 27, 28, 29] that the so-called z-score which is the rescaled quantity

$$z(p,Q) := \frac{W(p,Q) - \overline{W}_p}{\sigma_p},$$

where $\overline{W}_p$ and σ_p are the mean and standard deviation of $W(p,Q)$ when Q runs over all conformations in a database of known protein structures, can be used to identify the native fold P among a large set of possible conformations. Conversely, this z-score can be used as an approach to *inverse folding*: Given a fixed backbone conformation Q one could search for sequences p that give z-scores $z(p,Q)$ close to the z-score of the native sequence q. Of course, only structures that are already in the database can be searched for.

We have convincing evidence supporting this view. Markus Jaritz, personal communication, compared the z-scores of perturbed structures. These structures have been obtained by "heating" the structure in a molecular dynamics simulation and then cooling (and compactifying) it again. Figure 14 clearly shows a strong correlation between the rms-distance of the conformation from the wild-type and the z-score.

A second line of evidence comes from X-ray structures measured at different resolution by different labs (See table 2). In fact, the z-scores become better with increasing resolution of the structure determination.

Table 2. z-scores X-ray structures of the same protein at different resultions.

rms	z	z_{pair}	z_{surf}	Name
2.0	−13.56	−11.00	−9.37	1spa
2.4	−12.93	−10.18	−9.18	1aaw
2.5	−12.77	−9.91	−9.16	3aat
2.8	−11.39	−7.54	−9.14	2aat

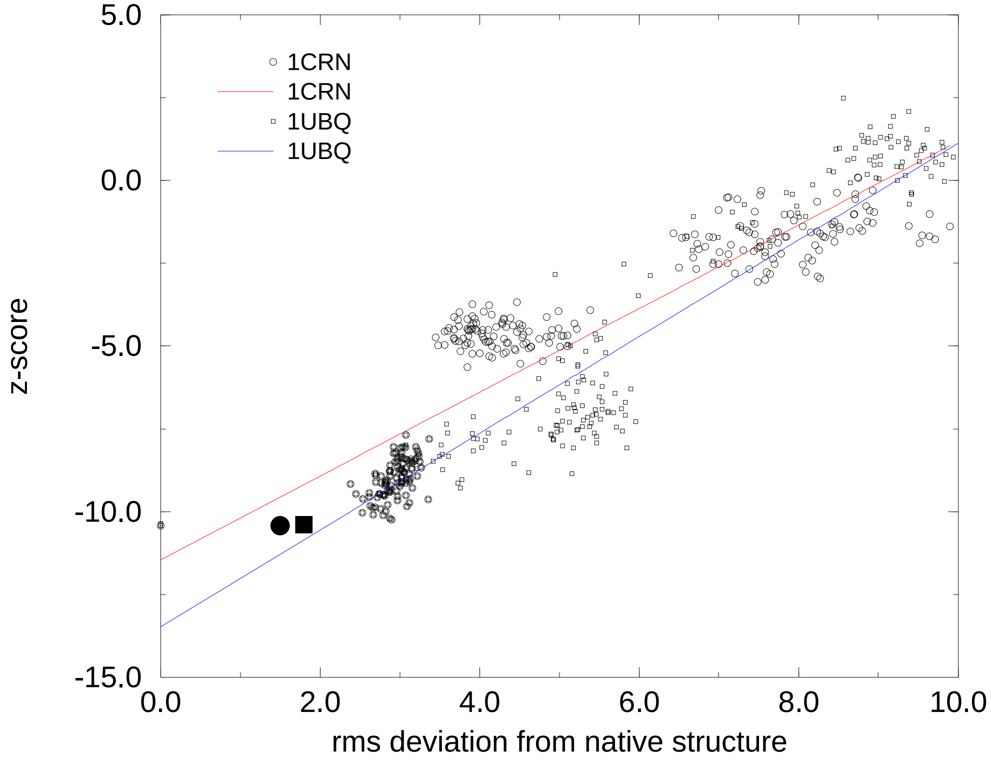

Figure 14: z-scores of protein structures obtained from perturbing them by high-temperature molecular dynamics simulations with cooling. The two big symbols show the native structures at their resolutions.

An alternative interpretation of the z-scores is as a measure on how close the native structure $S(q)$ is to the native structure $S(p) = P$. Hence we would use

$$d_P(q) := z(q, P) - z(p, P)$$

as an approximation to a structure distance between $P = S(p)$ and $S(q)$.
As a renormalized form of the z-scores one might use

$$\zeta(q, P) := \frac{z(q, P) - \langle z(x, P)\rangle_{x \in Q_\kappa^n}}{z(p, P) - \langle z(x, P)\rangle_{x \in Q_\kappa^n}}$$

This re normalization has the advantage that $\zeta \approx 1$ for perfectly fitting sequences and $\zeta \approx 0$ for random sequences. This rescaling should allow for some semi-quantitative comparison between sequences/structures of different chain length.

4.3. The Protein Structure Analysis Program (PROSA)

Several methods are available to decide whether or not a sequence is likely to fold into a given structure. For our studies we used the empirical mean force potentials developed by Sippl and coworkers (Center for Applied Molecular Engineering, University of Salzburg) [4, 17, 27, 28, 29] as implemented in the `PROSA` package. The potentials are derived from statistics of known three-dimensional protein structures and have been used successfully to identify a proteins native fold among a large set possible conformations. `Prosa II` is a powerful tool in protein structure research, it supports and guides studies aimed at the determination of a protein's native fold. It is helpful for experimental structure determinations and for modeling studies. Usually, the calculation of native folds of proteins from amino acid sequences is still impossible, even though modeling by homology has turned out to be quite successful in several cases. In general, if the native structure of a protein is needed there is still no escape from X-ray analysis and/or NMR-spectroscopy. Unfortunately, these techniques are time consuming and fail in many cases due to experimental problems (e.g. lack of isomorphous derivatives, size of the protein etc.), and the only remaining possibility is an attempt to build a model.
The `Prosa` program not only offers the possibility to evaluate experimentally determined protein structures, to identify incorrectly folded proteins (or sections of proteins), it is also a useful tool for the evaluation of theoretical models. Furthermore it is the basis of our approach to the inverse folding problem. The `Prosa` program based on the previously discussed knowledge based potentials of mean force allows us to study a large number of aspects of protein structures. It supports the following features and options:

`Prosa` is capable of reading both PDB and BBN (binary backbone) protein structure files. It has a `writebbn` command which can be used to generate BBN files if necessary. It can also read PDB files containing only C^α coordinates. This offers the option to analyze C^α traces. During the startup of the program C^β potentials are loaded per default, but it is possible to additionally load the C^α potentials and conduct calculations using both potentials.
The default polyprotein for the calculations of z-scores is *pII3.0.short.ply* which consists of 125 protein molecules. A larger protein (*pII3.0.long.ply*) made of 233 modules is provided as an option. These polyproteins are constructed from protein modules which are connected by linker regions. A polyprotein is a device for the

generation of alternative conformations for a given amino acid sequence. These conformations have a good stereo chemistry and have many features of native protein folds. The set of conformations derived from the polyprotein represents a sample of the conformation space ($\equiv$ shape space) of a given protein. The `Prosa` program can calculate a cost function (*z-score*) for a given protein conformation respectively for a given amino acid sequence on a certain protein structure.

The **z-score** is determined by the following method:

The amino acid sequence of the protein is combined with all conformations in the polyprotein and the energies of all conformations are calculated. The z-score is derived from the resulting energy distribution. The z-score z_p of the protein is obtained from the energy W_p of the protein by

$$z_p := \frac{W_p - \overline{W}}{\sigma_p}$$

where $\overline{W}$ is the average energy of all fragments derived from the polyprotein and σ_p is the associated standard deviation. The **total energy** is a combination of pair and surface energies. Pair interaction energies are calculated for residue pairs whose distance k along the sequence is $k_{lower} \leq k \leq k_{upper}$. The default values are $k_{lower} = 1$, $k_{upper} = 600$. However, if one is only interested in short range energy contributions (e.g. sequence separation $k \leq 9$) this variables can be reset. Pair energies are calculated in the distance range [*pot_lb, pot_ub*] Å. Outside this range energies are zero. *pot_lb* and *pot_ub* are variables that can be set depending on the energy contributions one is interested in (e.g if one is not interested in close contacts *pot_lb* can be increased etc.).

4.4. Adaptive Walks

For the "optimization" of our amino acid sequences we used the simplest possibility, an adaptive walk. In general, an adaptive walk will try a random mutation, and accept it if the cost function (in the present case the z-score) decreases. Here, a mutation means the exchange of one amino acid. If no advantageous mutation can be found, the procedure stops, and we may start again with a new initial string I_0. A disadvantage of the adaptive walk is that it could easily get stuck in a local optima (especially in case of the restricted alphabets). More elaborate optimization procedures (e.g. gradient walks) could avoid that, but in general they need more steps to find a solution. Even if several attempts are needed an adaptive walk therefore performs very well, except possibly for very rare structures. A typical adaptive walk employing the PROSA program is shown in Figure 15.

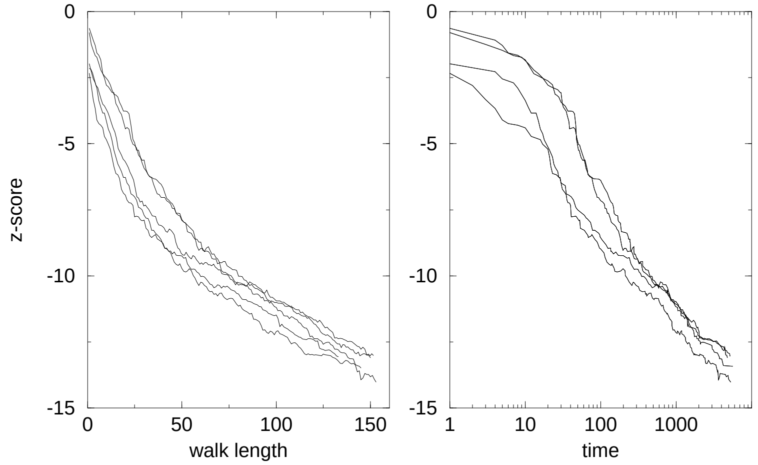

Figure 15: Adaptive walks on the 2TRX structure as a function of the z-score and as a function of the time.

The final sequences derived from the adaptive walks on the TRX structure had z-scores about 50% better than the score of the wildtype sequence (-9.22). This may seem surprising at first, however, there is no reason why the wild-types z-score should be optimal, too much structural stability might even be detrimental to the proteins function. On the other hand these results could be interpreted in respect of the accuracy of the potentials of mean force. The results depicted in Figure 14 show that the z-scores of the native structures improve with increasing resolution i.e., that the data contained in the resolved structures is by no means completely accurate. Therefore the fact that sequences with z-scores better than that of the native structures can be modeled could be an artifact of this inaccuracy. In case of the 2TRX thioredoxin we found that almost every 5th mutation of the wild-type sequence would improve the z-score. Although the scores are steadily improving even after ≈ 150 steps the time needed to find each additional sequence becomes very long at the end of the simulation. Continuing the simulation to z-scores way beyond those of native proteins, is clearly useless, as it would optimize the sequences with respect to the "noise" in the potentials.

One of the simplest ways to check whether the inverse folding generates plausible sequences and to see which regions of sequence space are explored during the inverse folding is to analyze the resulting distribution of amino acid frequencies. Figure 16 compares the amino acid frequencies from 9 inverse folded sequences to that of the wild-type sequence and the average composition of sequences in the SwissProt database. As can be seen the distribution is reasonably close to the expected mean composition. In other words, the inverse folding procedure explores the same regions of sequence space as typical proteins.

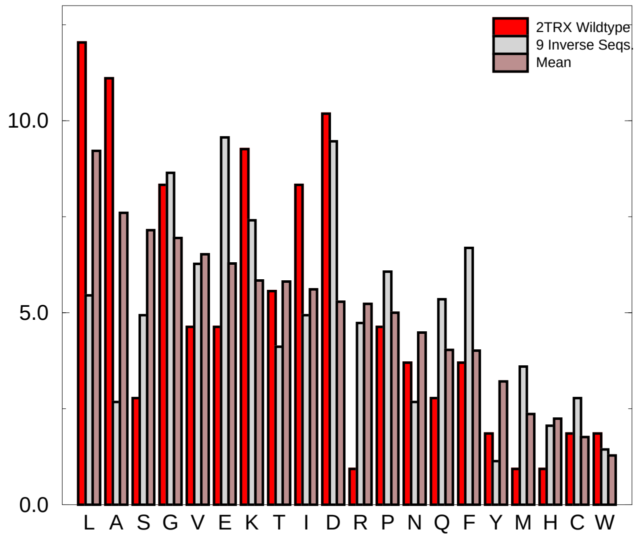

Figure 16: Mean amino acid composition of 9 inverse folded sequences compared to the wild-type composition and the mean over all sequences in the SwissProt database.

4.5. Neutral Networks

Our preliminary studies showed that the sequences folding into the same secondary structure S are randomly distributed in sequence space. Because of the high probability for finding neutral neighbors these sequences are not isolated, but form connected structures in sequence space. Hence, the question arises, how far such sets of neutral sequences extend. This can be tested in the following computer experiment. Starting from an initial sequence I_0 which we derive from a previously conducted adaptive walk, we construct a monotonously diverging "neutral path" by mutating our test sequence I_n, accepting the mutated sequence I_{n+1} if the mutation is neutral $\mathcal{S}(I) = \mathcal{S}(I_0)$ and the Hamming distance does not decrease $d(I_{(n+1)}, I_0) \geq d(I_n, I_0)$. As mutations we allow the exchange of a single amino acid in the reference structure.

The length $\mathcal{L}$ of a neutral path is the Hamming distance between the reference sequence and the last sequence, and hence a lower bound on the diameter of the connected "neutral network". Clearly, a neutral path cannot be longer than the chain length, $\mathcal{L} \leq n$.

The union of all neutral paths probably forms a dense neutral network, as in the case of RNA secondary structures. Of course, this need not be the case in general: rare structures may have short neutral paths confined to small disjoint regions in sequence space. Nevertheless, neutral nets are not a peculiarity of the few most frequent structures.

From the existence of such neutral networks one can expect far reaching consequences for evolutionary optimization where the fitness depends on structure [19]. Given a suitable error frequency, an evolving population should perform a random walk along the neutral net, until it reaches a point where a better secondary structure can be reached within a few mutations (i.e. a neutral net with higher fitness comes sufficiently close). During the times where the population diffuses on the neutral net, only the phenotype is conserved while genotypic information is unstable. For even lower error frequencies the population should localize in sequence space at a point on the neutral net where the number of neutral neighbors is especially large. Figure 17 shows the length of the neutral paths from computer experiments using the lysozyme structure.

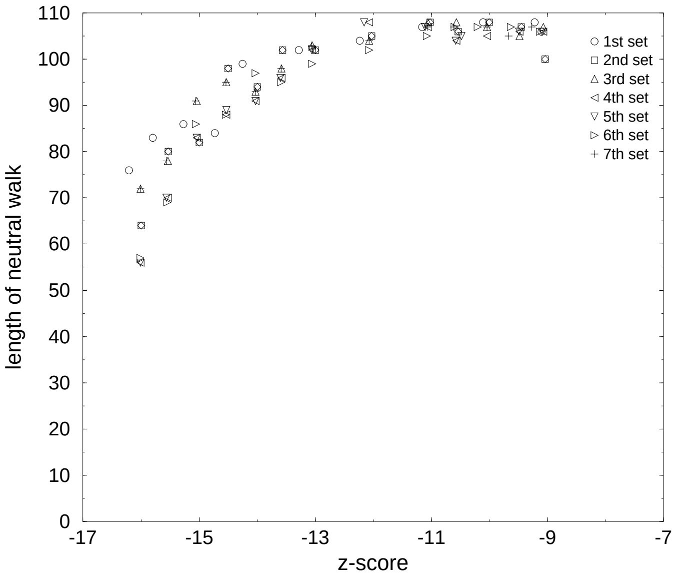

Figure 17: The length of neutral paths on the 1LYZ lysozyme structure at different z-scores

As long as the active site of an (enzymatic) protein is not touched, mutations along neutral paths and networks could be thought of as so called "silent mutations". This is the biological term for mutations that lead to the exchange of an amino acid in a protein without impairing its function. It can easily be understood that the existence of neutral neighbors for protein structures is essential for the survival of every organism in an environment that constantly provokes many different types of mutations.

4.6. Secondary Structures

Whether sequences predicted in the above described ways do indeed fold into the desired structure can ultimately only be answered by experiment. One way to test whether a computed sequence is plausible, is to try to predict its secondary structure and compare it to the known secondary structure of the target conformation. The best available algorithms combine secondary structure prediction with a search for homologous sequences and thereby attain accuracies over 70% for the assignment of residues to helix, strand and loop regions [24]. Since our inverse folded sequences have no or little homology to known sequences we have to expect somewhat lower accuracies.

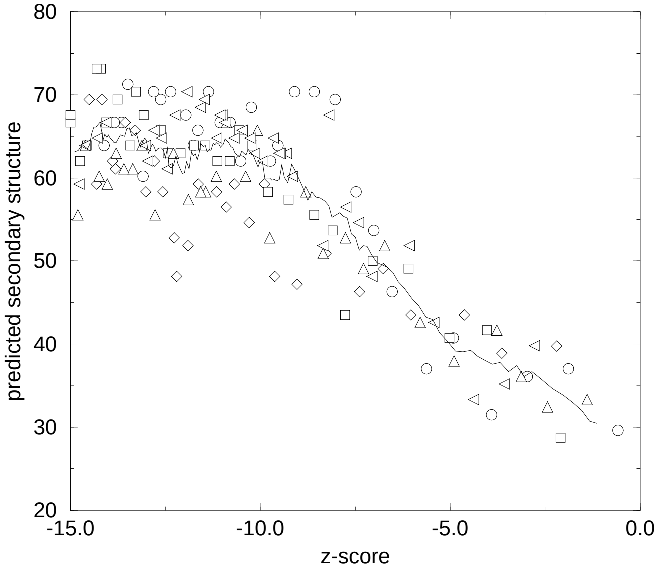

Figure 18: Identity between predicted and 2TRX secondary structure as a function of z-score for sequences from the 5 adaptive walks shown in 16. The full line is a running average of all 5 data sets.

We mostly used the program SOPM by Geourjon and Deleage [11]. It predicts ≈ 65% of residues in the 2TRX wildtype correctly. Similar results were achieved using the PHD method of Rost and Sanders [24]. Figure 18 shows the overlap between the 2TRX secondary structure and the SOPM prediction for every 5th sequence from the 5 adaptive walks shown in figure 15. The overlap between the predicted and 2TRX secondary structure at first increases with improving z-score, then saturates at about 65% once the z-score becomes better than that of the wild-type sequence, as expected. Note that the potentials depend only on distances between C^{α} or C^{β} atoms and surface exposure and make no use of secondary structure.

Sequences generated by the above procedure show little homology to the wild-type sequence or each other. The distribution of pairwise Hamming distances for 700 sequences with z-score ≈ -11 on the 2TRX structure can be seen in Figure 19. Although, they lie somewhat closer together than random sequences with a typical amino acid composition would (right curve), pairs with maximal distance still occur. The position of the maximum depends of course slightly on the z-score. The final sequences from 9 inverse foldings with z-scores better than -14 were identical to the wild-type in 16 out of 108 positions on average. This amount of homology is just about enough to be detected by a BLAST search. This suggests that sequences with similar structures are distributed widely and almost randomly over sequence space.

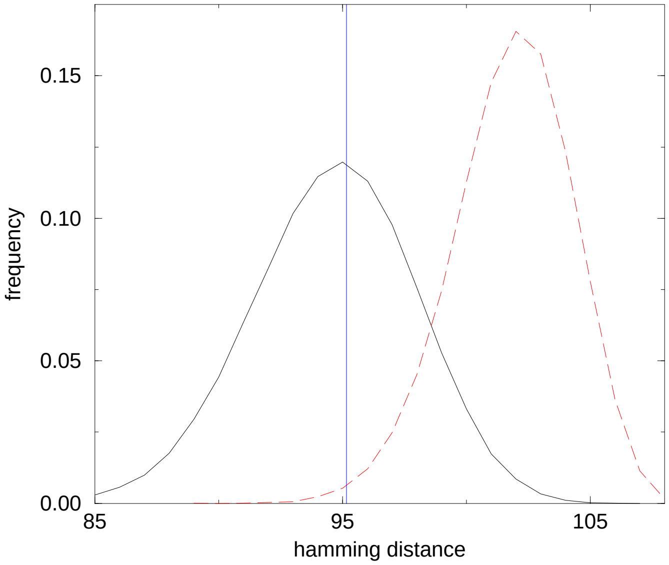

Figure 19: Distribution of pairwise hamming distances for 700 sequences designed to have z-scores ≈ -11 on the 2TRX structure (full curve) and for 500 random sequences of typical composition. The vertical line at 95.15 is the mean distance to the 2TRX sequence.

4.7. Substitution Patterns

We investigated whether or not the amino acids were substituted according to a discernible pattern during the adaptive and neutral walks. For this we screened a nuber of final sequences derived from these walks for hydrophobicity-hydrophility patterns. So far we have not found any obvious patterns, there seems to be no bias in the substitution frequency of the amino acids during the adaptive walks and during the search for neutral neighbors. Figure 20 shows a typical result.

```
-+++-+-++++-+++-+-++++---+--+++++-+++-++--++--++-+++-----+-+++-+++-+++++++--++---++++++++-+--+-+++--++-+++-+
+++-++-++++-+++-++++++-+++--++++++-++-+++-+++-++++++++-++-+-++-+++-++-+-+-+-+++--+++++++++++---++++++-++++--
+++-++++-+++++--+++++-+---+-+-+++++++-+-+-+++-++++++-++-++--+-++++-+-+++-+----+-+++-++++---++++-++++++++++-+
++--++-++++-+++-++-++-+--+----++++-++++---++++++-++++--++---++-++---++++++-+--++-++-+++++-+++++-+--++++++++-
+++-++---+++++--++++++--++-+++++++-+++-++-++--++++-+-++-+++-++-++-++++++++-++++--+++++-+--+-++++++-+++-++-+-
+-+-++-+-++--++-+++++-+------+++--+++--+--++++++++++---+-+++++++++--++-+++++-++--+++++++--+++-++++-++--+++-+
-+++++-++-+-+++-++++++---+-+-+-++-++++++---++--++-++++--++-+-++++++-++++++-+++---+-+++-+--++++-+++++++++++-+
+++-++++-++++++-++-++-------++++++-+++++++++-+++-++++--+--+-++-+++-+++++++-+-+--+--++++++--++-++++-++--++-++
++-++++++++++-+++-++-+++--++--++++-++++--+++++++++-+---++-+-+-++++-++-++++---+---++++--+-++++-+-+++++--++-+-
```

Figure 20: Hydrophobicity-Hydrophility patterns of sequences generated by neutral walks on the Thioredoxin structure at z-scores comparable to the wild-type

5. Experimental Data

It was our goal to study the extension of neutral paths and neutral networks for the four protein structures discussed in section 3 : Thioredoxin, Lysozyme, Crambin and Ubiquitin. Furthermore we explored the possibility of creating random sequences which would yield comparable or better z-scores than the wild-type. The last question we approached, was the number of amino acid types necessary to create sequences with z-scores equal or better than that of the wild-type. Of course, it was not possible to study all possible combinations amino acids, therefore we concentrated on the following alphabets of polar (P) and hydrophilic (H) amino acids: 3 alphabets of type HPHH ... ADLG, 1 alphabet of type HPH ... ADL, 1 alphabet of type PHP ... QLR and 3 alphabets of type HP ... LD, AS, LS. Recent experimental studies [5] showed that it is possible to create proteins with significant α-helical content and folded structures with native-like properties with the help of synthetic genes which encode random sequences of the amino acids QLR (glutamine, leucine, arginine). These proteins differ from natural proteins by their high resistance to denaturant-induced and thermally induced unfolding. These findings led to the inclusion of this combination of amino acids into our investigation of restricted alphabets. The restricted alphabet ADLG is also a good candidate for these studies as it is believed to be a “primordial” amino acid alphabet employed in the early stages of the evolution of the genetic code

5.1. Thioredoxin

As a first step, we conducted a number of adaptive walks on the Thioredoxin structure starting from random sequences. The structure we used to generate the binary backbone files was `2trxA.pdb`. From these adaptive walks we derived various sequences with z-scores equal or better than that of the wild-type (see Figure 21). The z-scores were calculated using both C_α and C_β-potentials i.e. sequences were accepted only if the mutation improved both the C_α and the C_β-potential.

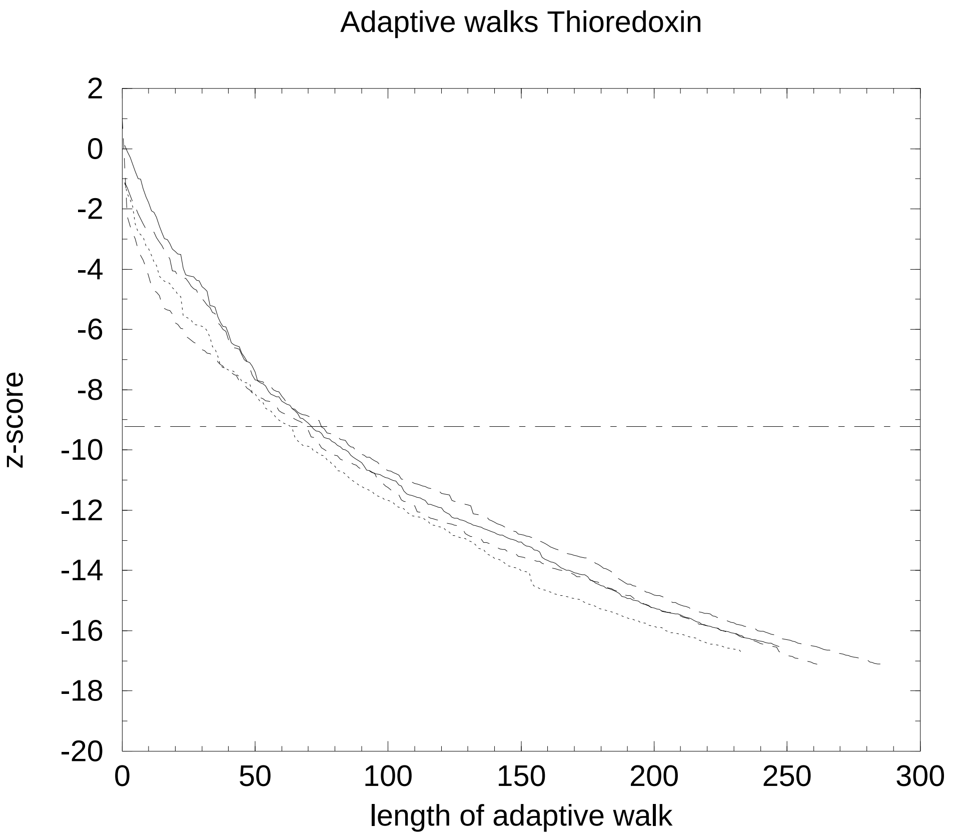

Figure 21: Results of four adaptive walks on the TRX structure. Each started from a different random sequence. The horizontal line indicates the z-score of the wild-type

The sequences derived from the adaptive walks were subsequently used to study the extension of neutral paths on the structure of Thioredoxin. For this we created binary backbone files using the backbone file of the wild-type and a sequence with the desired z-score. The sequences used were that of the wild-type and others within a z-score range of -9.04 to -16.00. The amino acids for the random substitutions were chosen according to their natural frequencies which we determined according to their frequency in the Swiss Prot Data Bank. The z-scores were calculated using both C_α and C_β-potentials. Figure 22 shows some results. We can see that the lengths of the neutral paths ($\equiv$ sequence length $-$ number of not mutated amino acids) are roughly equal to the length of the protein, at z-scores comparable to that of the wild-type (-9.22). Even at z-scores 5 to 6 standard deviations better than the wild-type z-score, the length of the neutral paths is still greater than three quarters of the length of the protein.

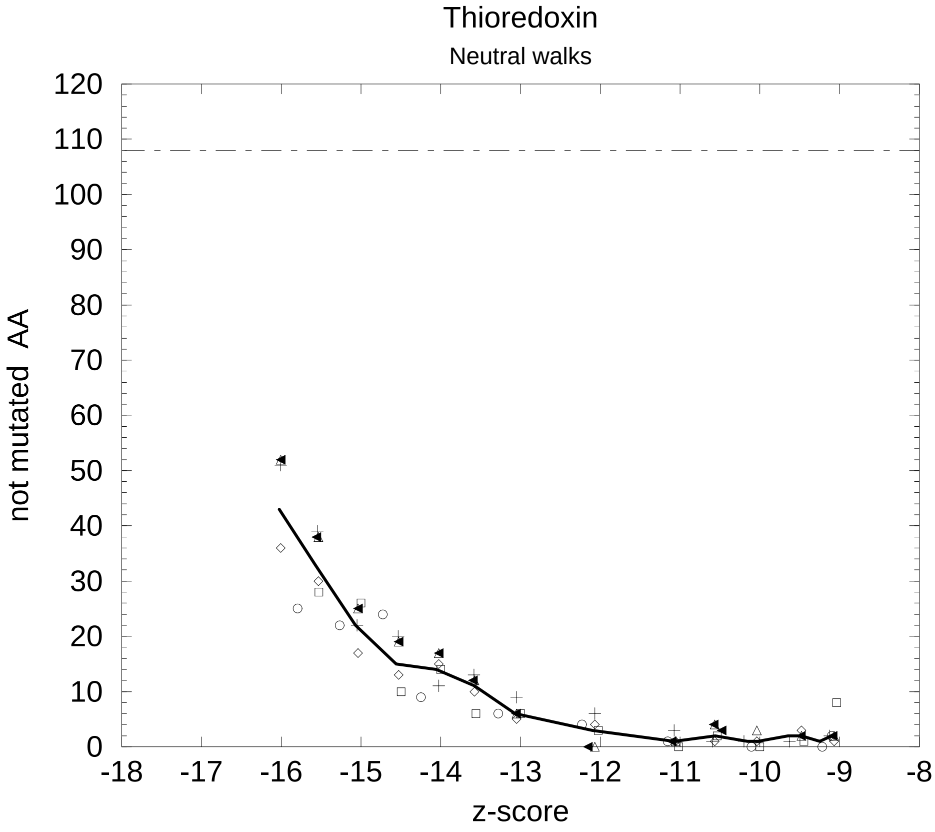

Figure 22: Results of different calculations of neutral paths on the TRX structure. The plot shows $\mathcal{L} - n$ as a function of the z-score. Dot-dashed line: Maximum number of amino acids in Thioredoxin (108). Solid line: Average number of not mutated amino acids.

Similar calculations, using restricted sets of amino acids were conducted. Figure 23 shows the results of two different sets of calculations which employed the alphabet ADLG. Random sequences were used to study the neutral paths. The z-scores were calculated using both C_α and C_β-potentials. We can observe that although the neutral paths are shorter than those using the entire alphabet, they still extend to 70% of the sequence length at z-scores comparable to that of the wild-type.

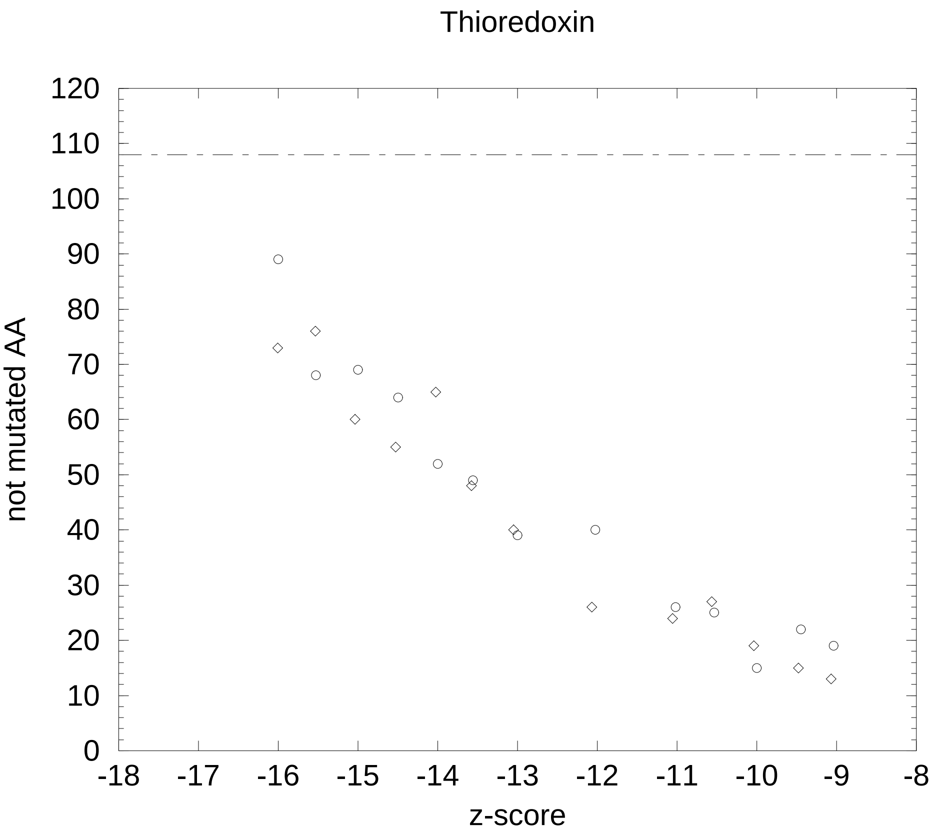

Figure 23: Results of two different sets of calculations of neutral paths on the TRX structure using only the amino acids ADLG. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of amino acids in Thioredoxin (108).

The question which arose next was, whether or not it is possible to generate sequences with z-scores comparable or better than that of the wild-type, with the above mentioned alphabets. Figure 24 shows the results of adaptive walks on the Thioredoxin structure, using the different restricted alphabets. The z-scores were calculated using both C_α and C_β-potentials. The results clearly show that, given the right combination of amino acids (e.g. AS, LS, ADL, ADLG) two to four amino acids are sufficient to create sequences with the desired z-scores. However it was not possible to create such sequences with the combination QLR which yielded interesting results in experimental studies [5].

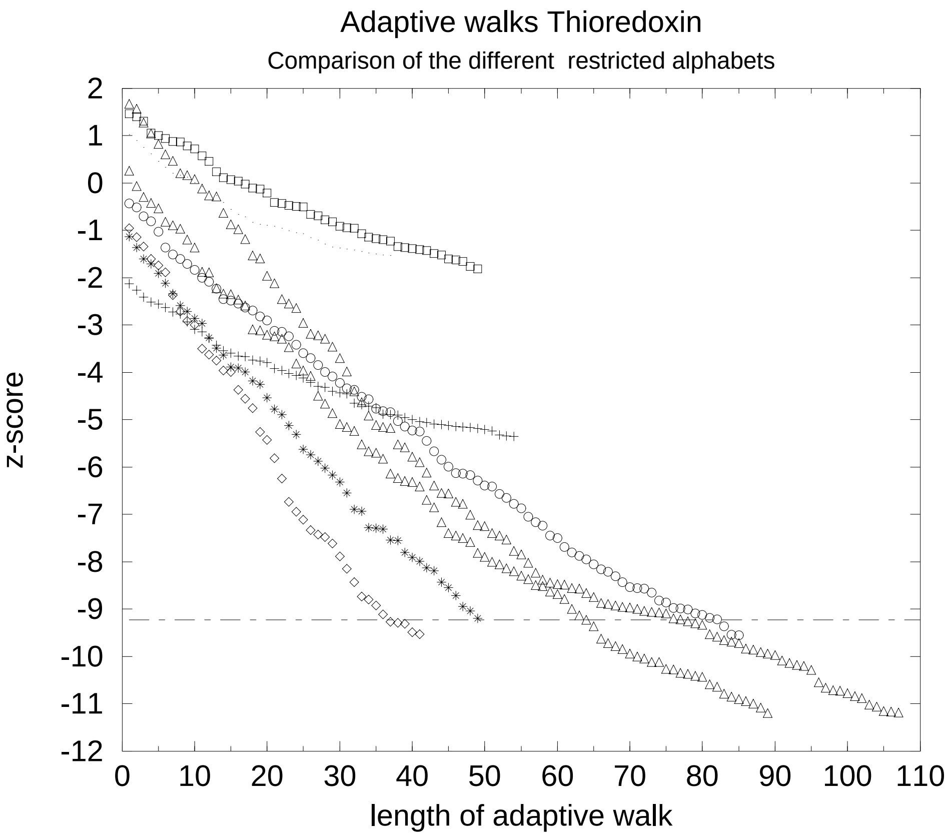

Figure 24: Comparison of the length of adaptive walks on the TRX structure using different sets of amino acids. Dot ... AD, Circle ... ADL, Square ... AS, Diamond ... DL, Plus ... QLR, Triangle up ... ADLG, Star ... LS. The horizontal line indicates the z-score of the wild-type.

Figure 25 shows the minimum z-scores achieved with the different combinations of amino acids. If we bear in mind, that the z-score of the wild-type is -9.22, we see that we have found at least four promising combinations of amino acids.

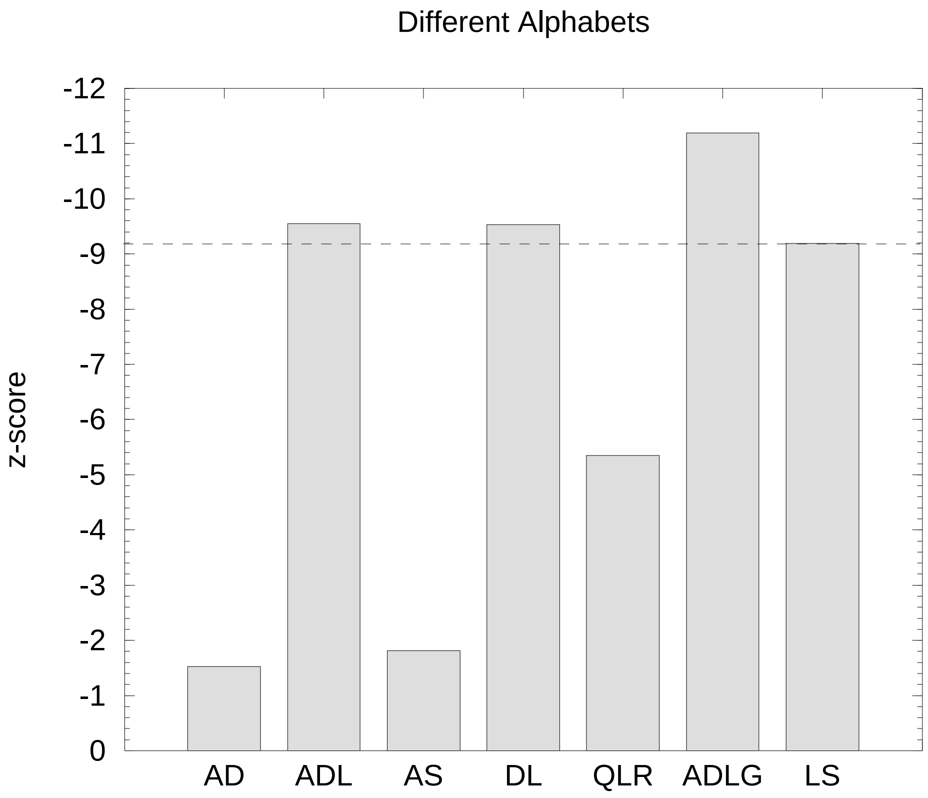

Figure 25: Minimum z-scores achieved by different adaptive walks on the TRX structure using different restricted alphabets.

Figure 26 shows the single steps of two adaptive walks conducted with alphabet 3 (DL). Obviously, comparably few random mutations are necessary to achieve the minimum z-score of -9.53 (respectively -9.09) of this adaptive walk. If we compare the number of steps necessary to the the length of the corresponding adaptive walk (= 41, see Figure 24), we see that approximately every second mutation is accepted. If we bear in mind that the adaptive walks started from random sequences, we can conclude that the distance from any sequence to a sequence which fits on a given structure is short, even if only a restricted set of amino acids is used.

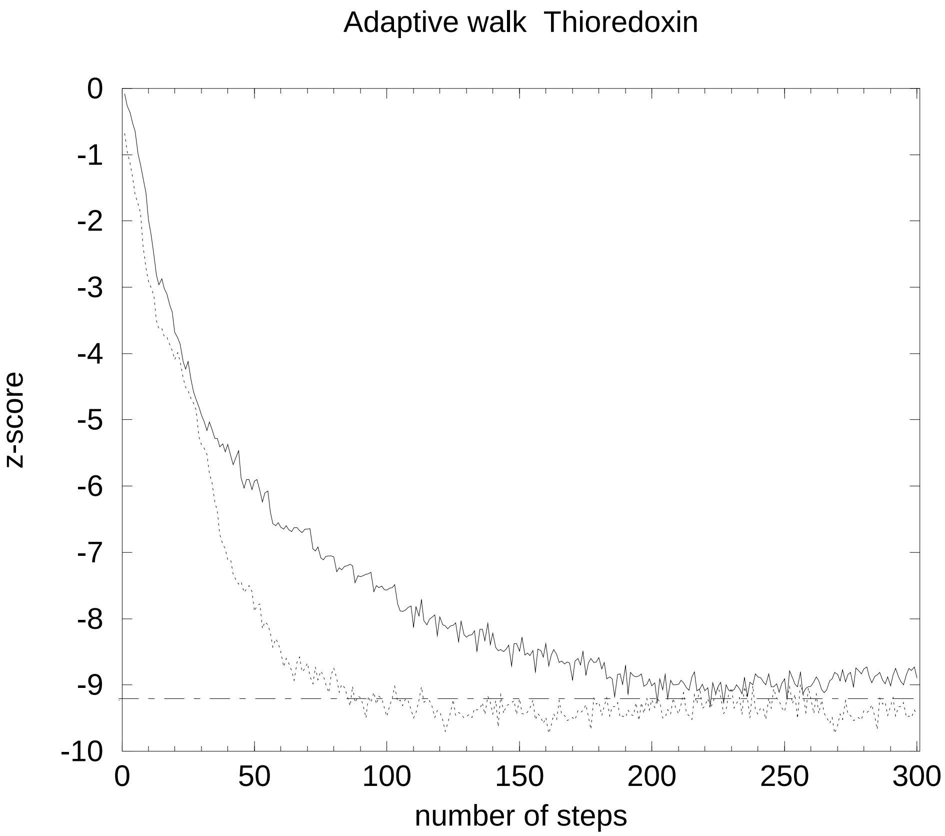

Figure 26: Two different adaptive walks on the TRX structure using the amino acids DL. Dot-dashed line: z-score of the wild-type.

5.2. Crambin

Again we performed a number of adaptive walks, in this case on the Crambin structure, starting from random sequences. The structure we used to generate the binary backbone files was `1cbn.pdb`. From these adaptive walks we derived various sequences with z-scores equal or better than that of the wild-type (see Figure 27). The z-scores were calculated using both C_α and C_β-potentials i.e. sequences were accepted only if the mutation improved both the C_α and the C_β-potential.

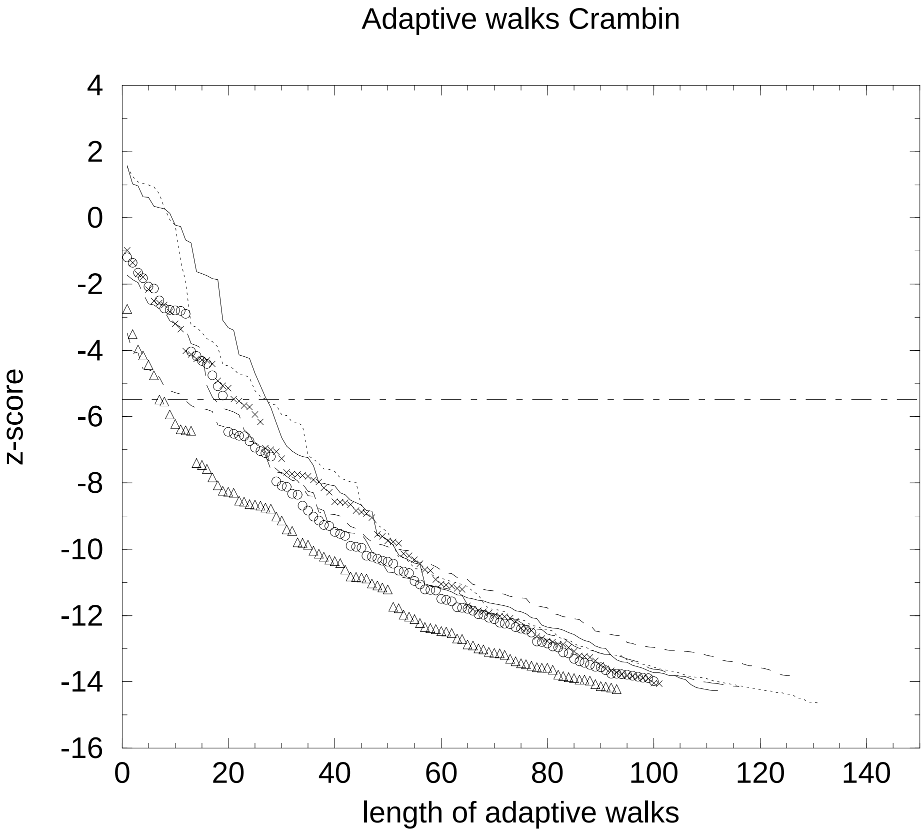

Figure 27: Results of six adaptive walks on the CBN structure. Each started from a different random sequence. The horizontal line indicates the z-score of the wild-type

The sequences derived from the adaptive walks were subsequently used to study the extension of neutral paths on the 1CBN Crambin structure. For this we created binary backbone files using the backbone file of the wild-type and a sequence with the desired z-score. The sequences used were that of the wild-type and others within a z-score range of -5.5 to -13.0. The amino acids for the random substitutions were chosen according to their natural frequencies which we determined according to their frequency in the Swiss Prot Data Bank. The z-scores were calculated using both C_α and C_β-potentials. Figure 28 shows some results. We can see that the lengths of the neutral paths are roughly equal to the length of the protein, at z-scores comparable to that of the wild-type (-5.5). Even at z-scores 6 to 8 standard deviations better than the wild-type z-score, the length of the neutral paths is still approximately three quarters the length of the protein.

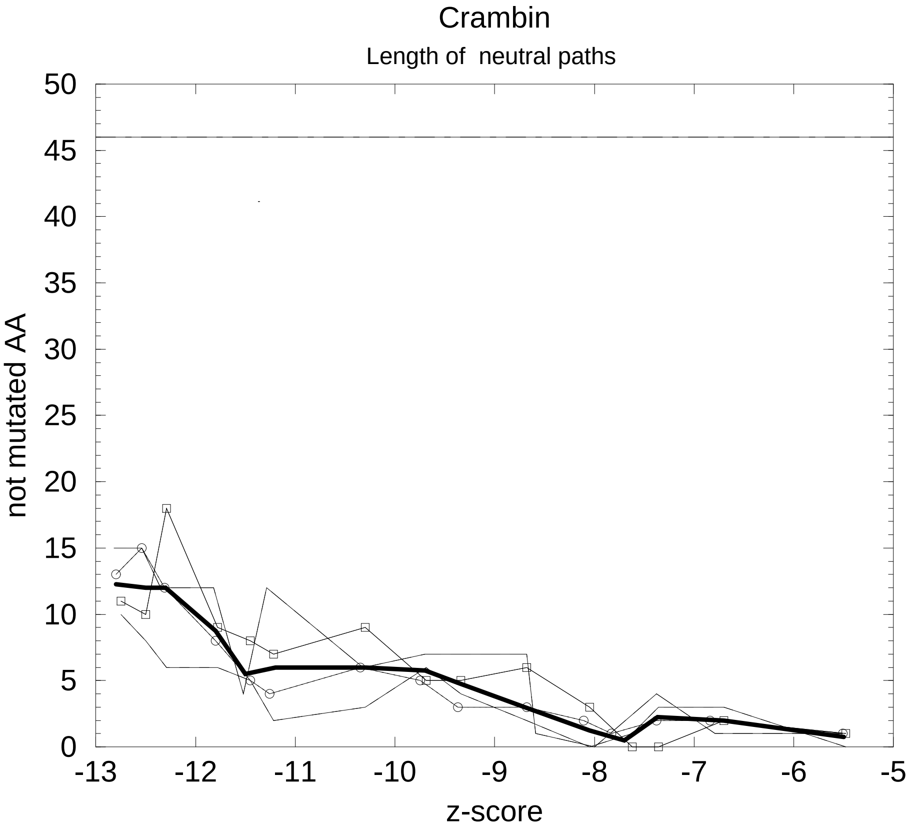

Figure 28: Results of different calculations of neutral paths on the CBN structure. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of amino acids in Crambin (46). Thick line: Average number of not mutated amino acids.

Similar calculations, using restricted sets of amino acids were conducted. Figure 29 shows the results of two different sets of calculations which employed the alphabet ADLG. Random sequences were used to study the neutral paths. The z-scores were calculated using both C_α and C_β-potentials. We can observe that although the neutral paths are shorter than those using the entire alphabet, they still extend to 90% of the sequence length at z-scores comparable to that of the wild-type and 2 to 3 standard deviations better.

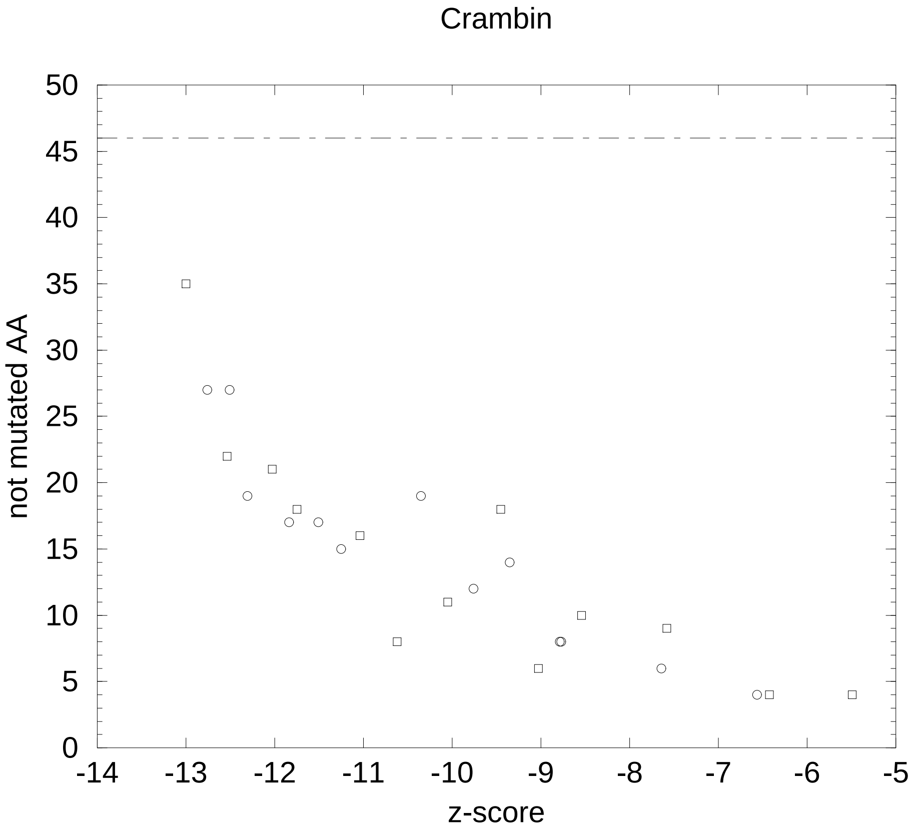

Figure 29: Results of two different sets of calculations of neutral paths on the CBN structure using only the amino acids ADLG. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of amino-acids in Crambin (46).

The next question was, whether or not it is possible to generate sequences with z-scores comparable or better than that of the wild type, using restricted alphabets. Figure 30 shows the results of adaptive walks on the 1CBN Crambin structure, using different restricted alphabets. The z-scores were calculated using both C_α and C_β-potentials. The results clearly show that, given the right combination of amino acids (e.g. AS, LS, ADL, ADLG) two to four amino acids are sufficient to create sequences with the desired z-scores. Again it was not possible to create such sequences with the combination QLR which yielded interesting results in experimental studies [5].

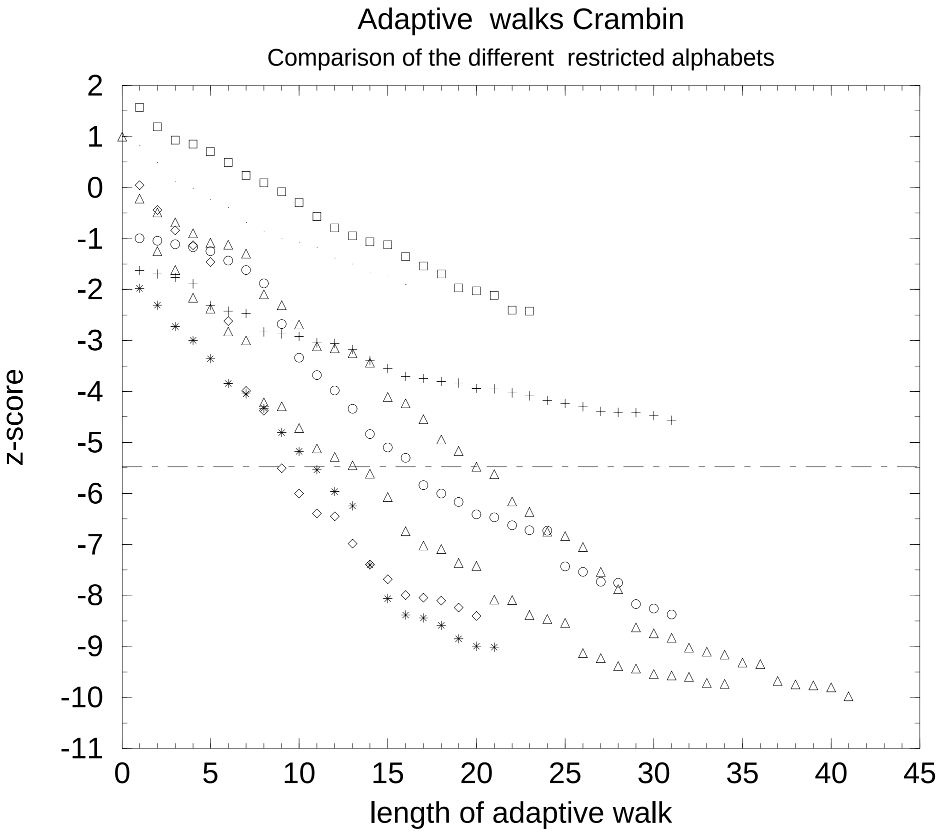

Figure 30: Comparison of the length of adaptive walks on the CBN structure using different sets of amino acids. Dot ... AD, Circle ... ADL, Square ... AS, Diamond ... DL, Plus ... QLR, Triangle up ... ADLG, Star ... LS. The horizontal line indicates the z-score of the wild-type.

Figure 31 shows the minimum z-scores achieved with the different combinations of amino acids. If we bear in mind, that the z-score of the wild-type is -5.5, we see that we have again found the same four promising combinations of amino acids.

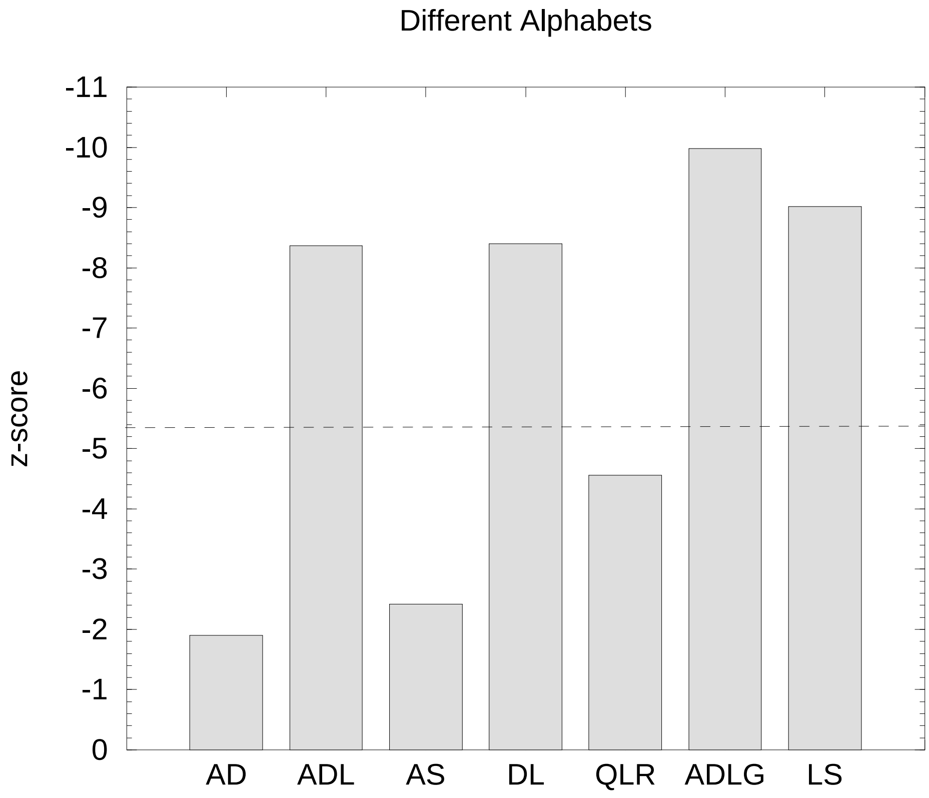

Figure 31: Minimum z-scores achieved by different adaptive walks on the CBN structure using different restricted alphabets.

5.3. Ubiquitin

Using the same procedures as mentioned earlier, we again conducted a number of adaptive walks, presently on the 1UBQ Ubiquitin structure, starting from random sequences. The structure we used to generate the binary backbone files was `1ubq.pdb`. From these adaptive walks we derived various sequences with z-scores equal or better than that of the wildtype (see Figure 32). The z-scores were calculated using both C_α and C_β-potentials i.e. sequences were accepted only if the mutation improved both the C_α and the C_β-potential.

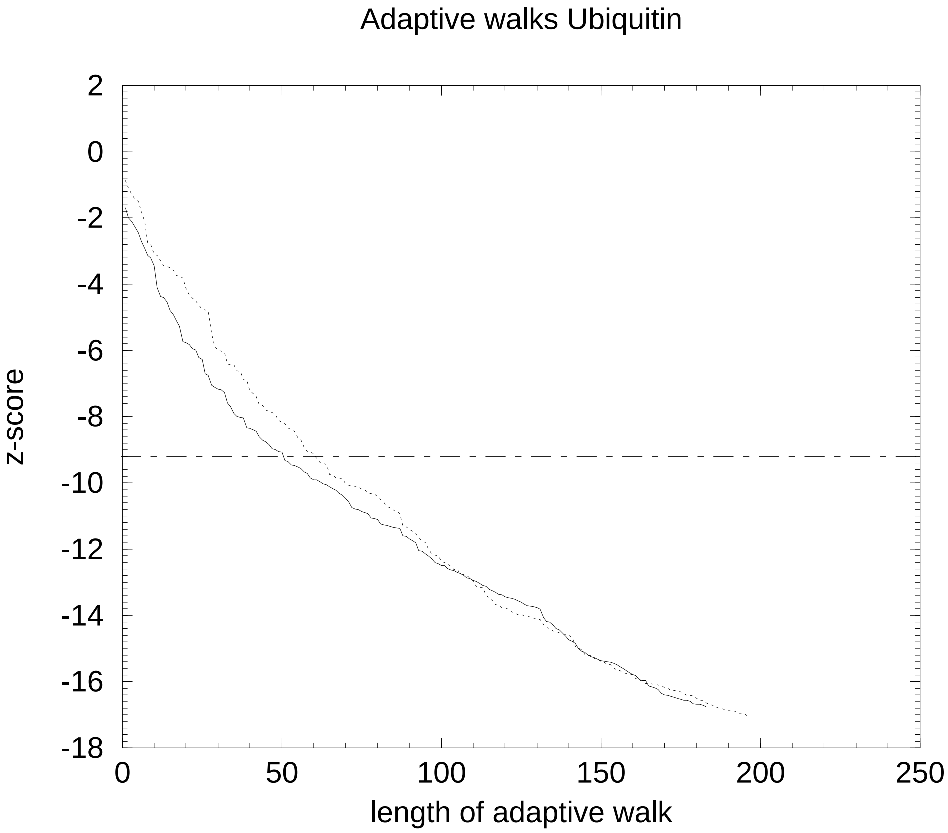

Figure 32: Results of two adaptive walks on the UBQ structure. Each started from a different random sequence. The horizontal line indicates the z-score of the wild-type

The sequences derived from the adaptive walks were subsequently used to study the extension of neutral paths on the 1UBQ Ubiquitin structure. For this we created binary backbone files using the backbone file of the wildtype and a sequence with the desired z-score. The sequences used were that of the wildtype and others within a z-score range of -9.3 to -17.0. The amino acids for the random substitutions were chosen according to their natural frequencies which we determined according to their frequency in the Swiss Prot Data Bank. The z-scores were calculated using both C_α and C_β-potentials. Figure 33 shows some results. Again we see that the lengths of the neutral paths are roughly equal to the length of the protein, at z-scores comparable to that of the wildtype (-9.3). At z-scores up to 7 standard deviations better than the wildtype z-score, the length of the neutral paths is greater than 60% of the sequence length.

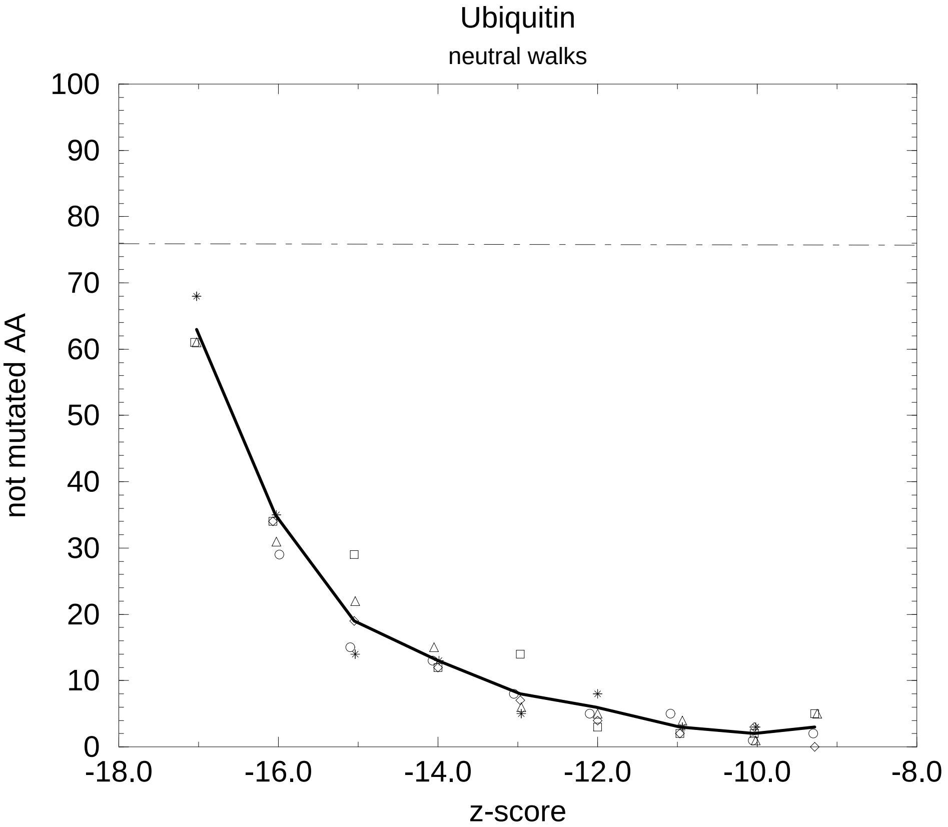

Figure 33: Results of different calculations of neutral paths on the UBQ structure. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of aminoacids in Ubiquitin (76). Thick line: Average number of not mutated amino acids.

Similar calculations, using restricted sets of amino acids were conducted. Figure 34 shows the results of two different sets of calculations which employed the alphabet ADLG. Random sequences were used to study the neutral paths. The z-scores were calculated using both C_α and C_β-potentials. We can observe that although the neutral paths are shorter than those using the entire alphabet, they still extend to 60% of the sequence length at z-scores comparable to that of the wildtype and slightly better.

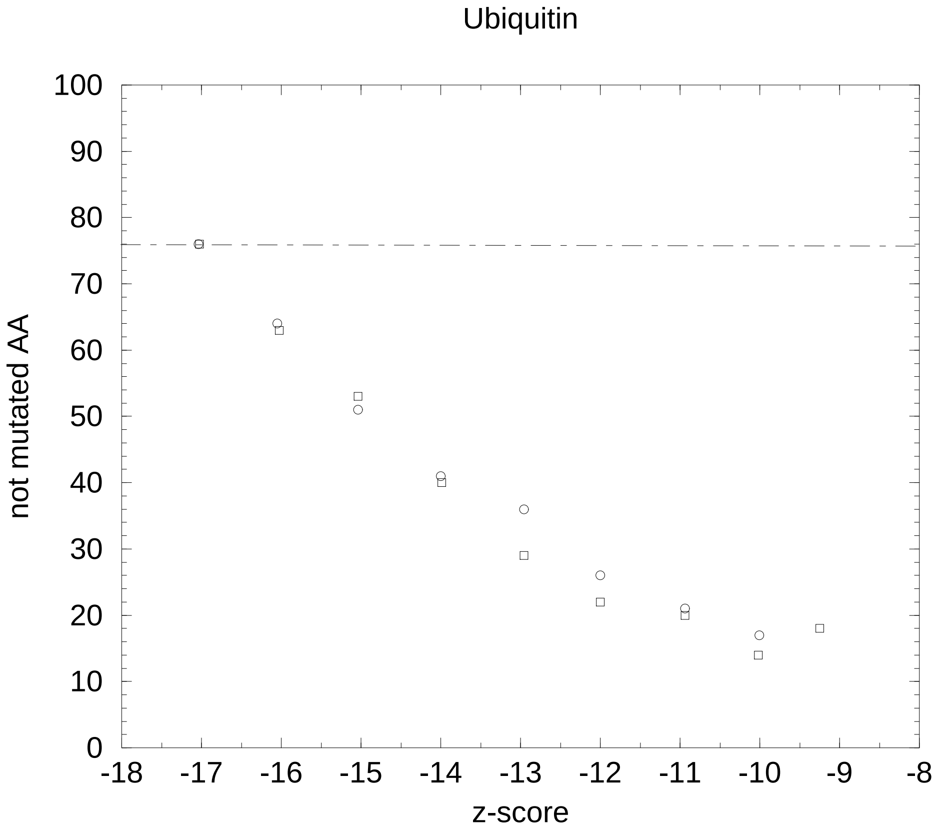

Figure 34: Results of different sets of calculations of neutral paths on the UBQ structure using only the amino acids ADLG. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of aminoacids in Ubiquitin (76).

Again we tried to answer the question, whether or not it is possible to generate sequences with z-scores comparable or better than that of the wildtype, using retricted alphabets. Figure 35 shows the results of adaptive walks on the 1UBQ Ubiquitin structure, using different restricted alphabets. The z-scores were calculated using both C_α and C_β-potentials. The results show that, given the right combination of amino acids (e.g. AS, LS, ADL, ADLG) two to four amino acids are sufficient to create sequences with the desired z-scores. Again it was not possible to create such sequences with the combination QLR which yielded interesting results in experimental studies [5].

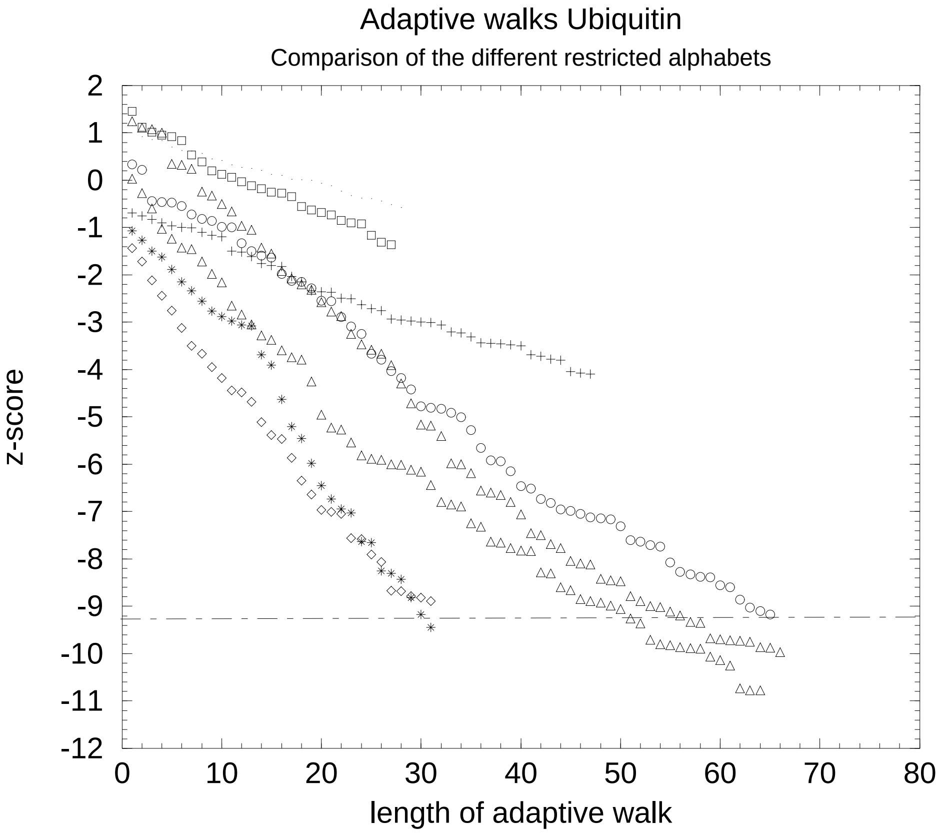

Figure 35: Comparison of the length of adaptive walks on the UBQ structure using different sets of amino acids. Dot ... AD, Circle ... ADL, Square ... AS, Diamond ... DL, Plus ... QLR, Triangle up ... ADLG, Star ... LS. The horizontal line indicates the z-score of the wild-type.

Figure 36 shows the minimum z-scores achieved with the different combinations of amino acids. The z-score of the wildtype is -9.3, we see that we have again found the same four promising combinations of amino acids.

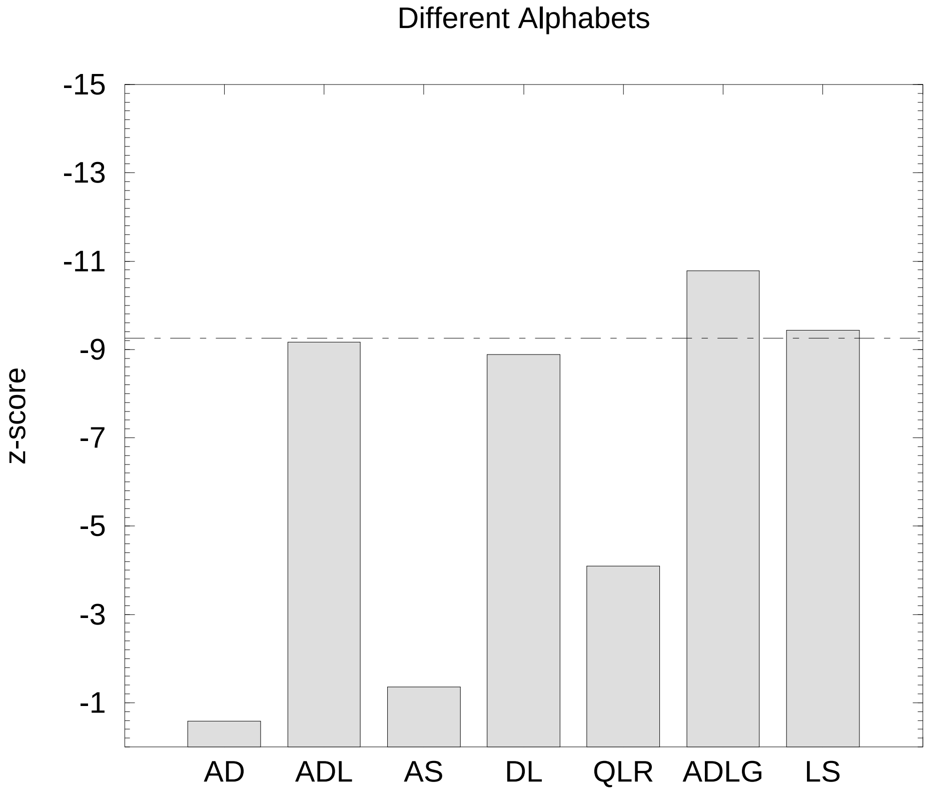

Figure 36: Minimum z-scores achieved by different adaptive walks on the UBQ structure using different restricted alphabets.

5.4. Lysozyme

Again we conducted a number of adaptive walks, in this case on the 1LYZ lysozyme structure, starting from random sequences. The structure we used to generate the binary backbone files was `1lyz.pdb`. From these adaptive walks we derived various sequences with z-scores equal or better than that of the wild-type (see Figure 37). The z-scores were calculated using both C_α and C_β-potentials i.e. sequences were accepted only if the mutation improved both the C_α and the C_β-potential.

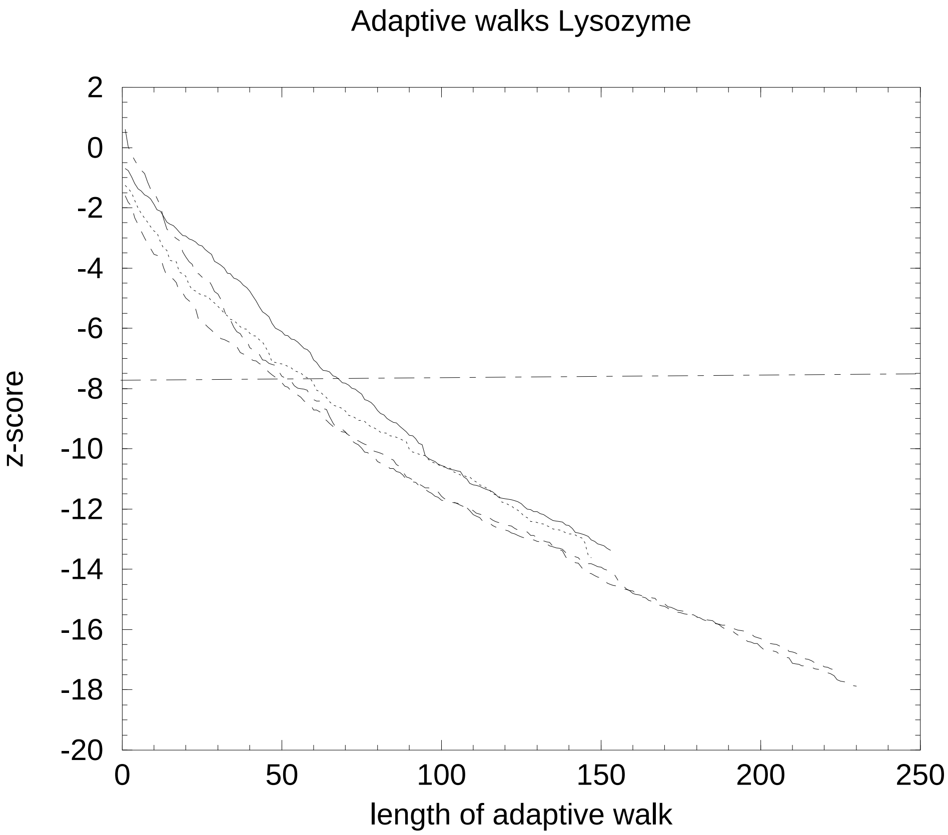

Figure 37: Results of four adaptive walks on the LYZ structure. Each started from a different random sequence. The horizontal line indicates the z-score of the wild-type

The sequences derived from the adaptive walks were subsequently used to study the extension of neutral paths on the 1LYZ Lysozyme structure. For this we created binary backbone files using the backbone file of the wild-type and a sequence with the desired z-score. The sequences used were that of the wild-type and others within a z-score range of -8.0 to -17.0. The amino acids for the random substitutions were chosen according to their natural frequencies which we determined according to their frequency in the Swiss Prot Data Bank. The z-scores were calculated using both C_α and C_β-potentials. Figure 38 shows some results. We can see that the lengths of the neutral paths are nearly equal to the length of the protein, at z-scores comparable to that of the wild-type (-7.7). At all z-scores, even those 9 standard deviations better than the wild-type z-score, the length of the neutral paths is greater than 60%of the length of the protein.

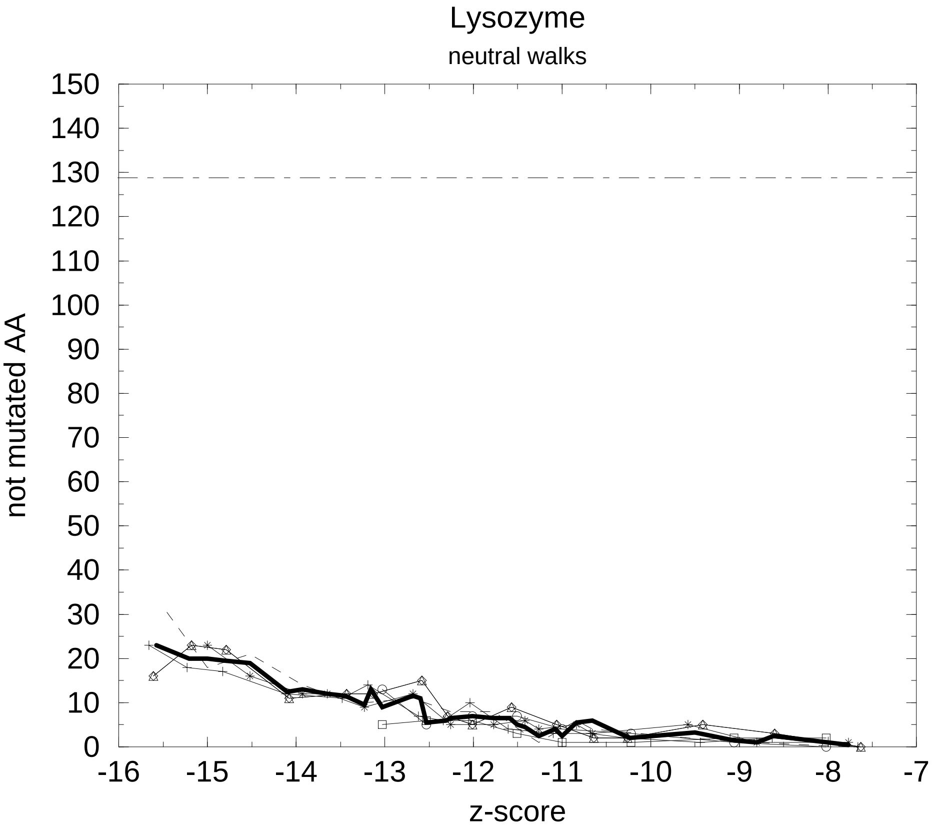

Figure 38: Results of different calculations of neutral paths on the LYZ structure. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of amino acids in Lysozyme (129). Thick line: Average number of not mutated amino acids.

Similar calculations, using restricted sets of amino acids were conducted. Figure 39 shows the results of two different sets of calculations which employed the alphabet ADLG. Random sequences were used to study the neutral paths. The z-scores were calculated using both C_α and C_β-potentials. We can observe that although the neutral paths are shorter than those using the entire alphabet, they still extend to 80% of the sequence length at z-scores comparable to that of the wild-type and 2 to 3 standard deviations better.

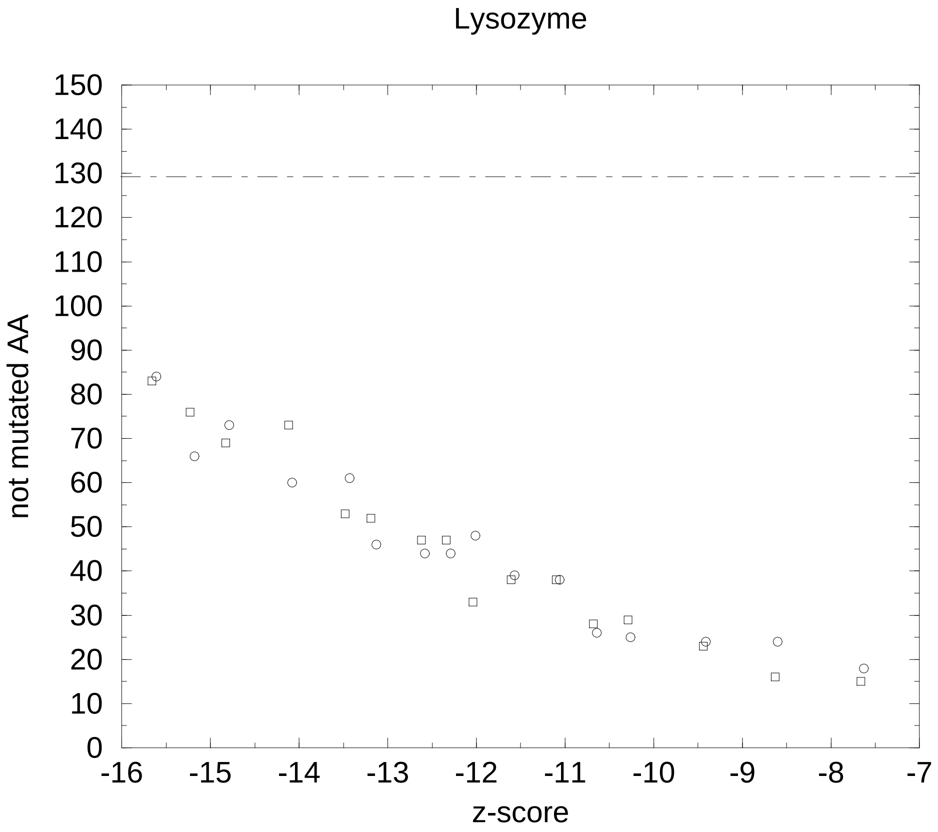

Figure 39: Results of two different sets of calculations of neutral paths on the LYZ structure using only the amino acids ADLG. The plot shows $\mathcal{L} - n$ as a function of the z-score. The horizontal line indicates the maximum number of amino acids in Lysozyme (129).

The next question was, whether or not it is possible to generate sequences with z-scores comparable or better than that of the wild-type, using restricted alphabets. Figure 40 shows the results of adaptive walks on the 1LYZ Lysozyme structure, using different restricted alphabets. The z-scores were calculated using both C_α and C_β-potentials. The results clearly show that, given the right combination of amino acids (e.g. AS, LS, ADL, ADLG) two to four amino acids are sufficient to create sequences with the desired z-scores. Again it was not possible to create such sequences with the combination QLR.

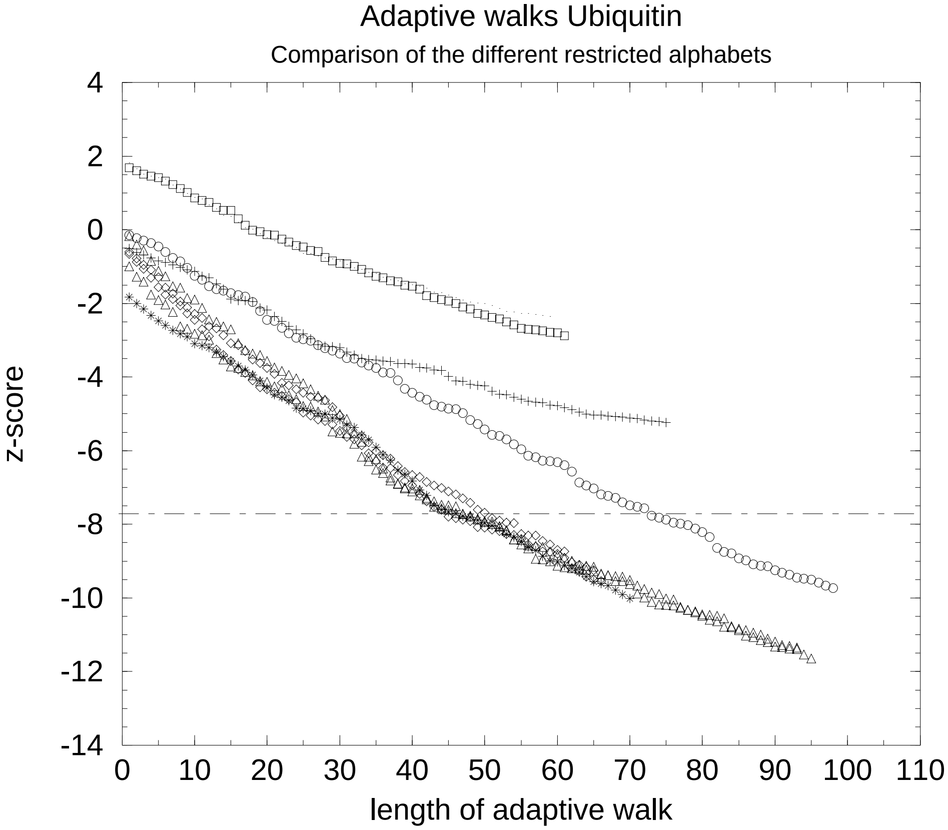

Figure 40: Comparison of the length of adaptive walks on the LYZ structure using different sets of amino acids. Dot ... AD, Circle ... ADL, Square ... AS, Diamond ... DL, Plus ... QLR, Triangle up ... ADLG, Star ... LS. The horizontal line indicates the z-score of the wild-type.

Figure 41 shows the minimum z-scores achieved with the different combinations of amino acids. We have again found the same four promising combinations of amino acids.

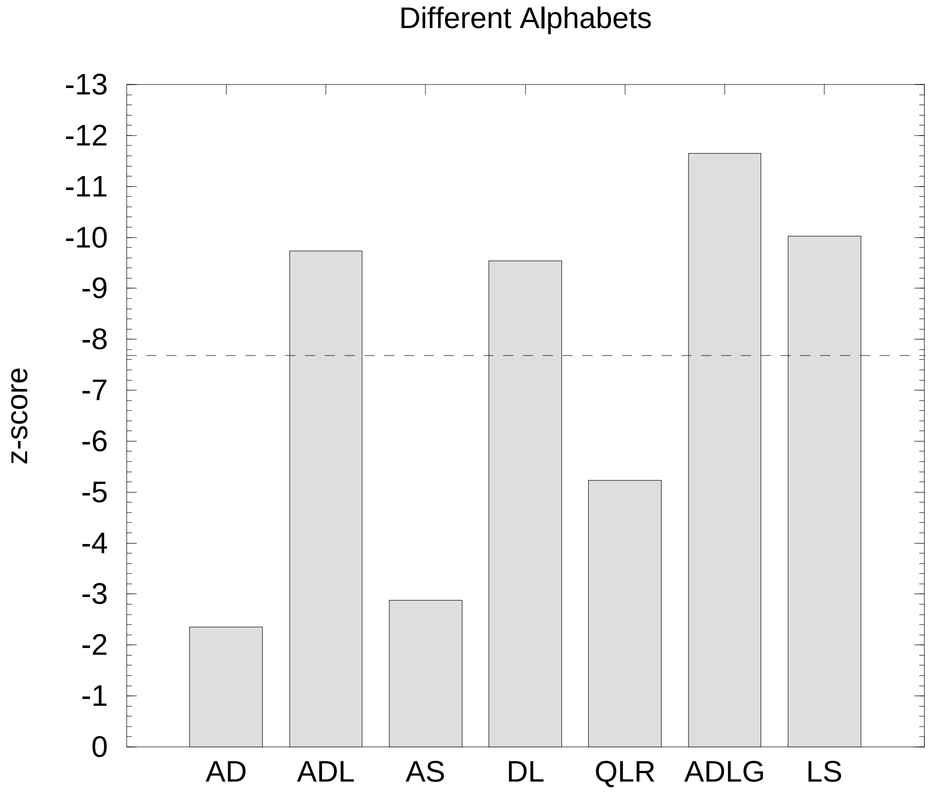

Figure 41: Minimum z-scores achieved by different adaptive walks on the LYZ structure using different restricted alphabets.

6. Conclusions and Outlook

The approach to inverse folding outlined in this work may be useful as a tool for studying general sequence structure relations in proteins as well as for protein engineering in biotechnological applications. We have found that sequences with z-scores as good or better than the wild-type score can indeed be found using a simple adaptive walk in sequence space, i.e., a random point mutation is introduced and accepted if and only if the z-score is decreased. The reason why it is not necessary to use a more sophisticated optimization technique is that local minima in the high dimensional sequence space are rare. It remains to be shown that the sequences we have found by adaptive and neutral walks do indeed fold into the target conformation.

We found that there is essentially no homology between the inverse folded sequences, the distribution of the amino acids is random-like. The neutral paths on protein structures extend to the length of the amino acid sequence at z-scores comparable to the wild-type score and better i.e., the neutral walks on all protein structures are very long, this is a strong indication for the existence of *Neutral Networks*. In regard to this aspect, the Sequence − Structure relations of proteins (see Figure 42) seems to be similar to the Sequence − Structure relations (see Figure 1). Consequently, we must pose the question, whether or not there is *Shape Space Covering* as in the RNA case [26, 15, 16]. For any evolutionary optimization it is of prime importance how big a volume in sequence space has to be searched in order to find a sequence with the desired properties. We may therefore pose the question how close to some given starting sequence a preselected secondary structure can be found. Stated differently the question is what radius a ball in sequence space must have to contain most common structures. This radius is called the shape space covering radius h_c. However, as the answer to this question would be computationally expensive in case of protein structures, we have not yet begun to tackle it.

Some of the investigated reduced alphabets show the same properties as the full set of amino acids, other don't. The results clearly show that the distinction between hydrophobic and hydrophilic amino acids is not sufficient to explain the differences between the various reduced alphabets considered in this work. Further investigations will be necessary to elucidate the effect of amino acid composition.

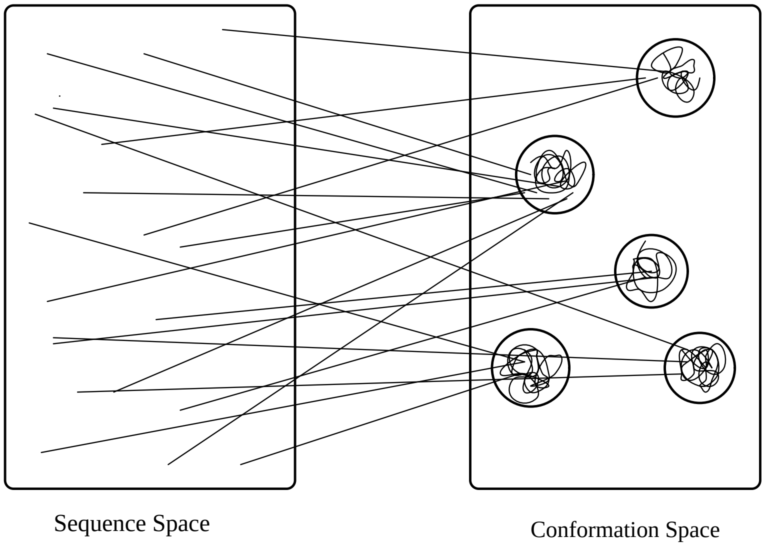

Figure 42: Sequence − Structure relation in the case of protein secondary structures: Sequences folding into a particular structure can be found anywhere in sequence space. Such sequences can be connected by extended nets of structurally neutral neighbors.

For small proteins the procedure we used is fast enough to produce large numbers of candidates, which could then be filtered using additional criteria such as other potentials, secondary structure prediction, analysis of hydrophobicity and packing along the chain to selected only the most promising ones. The accuracy of the potentials therefore need not be perfect. Whether it is sufficient can only be decided by experiment.

7. References

[1] A. Bauer and A. Beyer. An improved pair potential to recognize native protein folds. *Proteins*, 18:254–261, 1994.

[2] J. U. Bowie, N. D. Clarke, and C. O. Pabo. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, 7:257, 1990.

[3] J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164, 1991.

[4] G. Casari and M. J. Sippl. Structure-derived hydrophobic potentials — hydrophobic potentials derived from x-ray structures of globular proteins is able to indentify native folds. *J.Mol.Biol.*, 224:725–732, 1992.

[5] A. R. Davidson and R. T. Sauer. Folded proteins occur frequently in libraries of ramdom amino acid sequences. *Proc. Natl. Acad. Sci., USA*, 91:2146–2150, 1994.

[6] J. D.Bryngelson. When is a potential accurate enough for structure prediction? theory and application to a random heteropolymer model of protein folding. *The Journal of Chemical Physics*, 100:6038, 1994.

[7] W. Fontana, T. Griesmacher, W. Schnabl, P. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures. *Monatshefte der Chemie*, 122:795–819, 1991.

[8] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[9] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083 – 2099, 1993.

[10] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[11] C. Geourjon and G. Deleage. SOPM : a self optimised prediction method for protein secondary structure prediction. *Protein Engineering*, 7:157–164, 1994.

[12] A. Godzik, A. Kolzinski, and J.Skolnik. A topology fingerprint approach to the inverse protein folding problem. *J.Mol.Biol.*, 227:227–238, 1992.

[13] R. Goldstein, Z. Luthey-Schulten, and P. Wolynes. Protein tertiary structure recognition using optimized hamiltonians with local interaction. *Proc. Natl. Acad. Sci., USA*, 89:9029–9033, 1992.

[14] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. *Ismb*, 3:154–61, 1995.

[15] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, page in press, 1996. SFI preprint 95-10-099.

[16] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath. Chem.*, page in press, 1996. SFI preprint 95-10-099.

[17] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models — the calculation of low energy conformations from potentials of mean force. *J.Mol.Biol.*, 216:167–180, 1990.

[18] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125(2):167–188, 1994.

[19] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[20] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M.Babik, and M. H.Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685, 1993.

[21] C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65:44 – 45, 1968.

[22] R. Luthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83, 1992.

[23] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[24] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

[25] R. T. Sauer. Protein folding from a combinatorial perspective. *Folding & Design*, 1:R27–R29, 1996.

[26] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[27] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force — an approach to the knowledge-based prediction of local structures in globular proteins. *J.Mol.Biol.*, 213:859–883, 1990.

[28] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J.Computer-Aided Molec.Design*, 7:473–501, 1993.

[29] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993. URL: `http://lore.came.sbg.ac.at/Extern/software/Prosa/prosa.html`.

[30] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46:591–621, 1984.

[31] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[32] M. Zvelebil, G. Barton, W. Taylor, and M. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J.Mol.Biol.*, 195:957–961, 1987.

Table of Contents

Curriculum vitae

Aderonke Babajide

∗ 1966–06–03

1972–1974	:	Elementary School, 16th District, Vienna, Austria
1974–1985	:	German School Lagos, Nigeria
1985	:	Abitur at the German School Lagos
1985–1996	:	Studies of Biochemistry at the University of Vienna
2/1995–4/1996	:	Diploma thesis with Doz. Dr. Peter Stadler at the Institute of Theoretical Chemistry, University of Vienna