Neutral Networks in Protein Space

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Eingereicht an der
Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

von
Mag. Aderonke Babajide

Institut für Theoretische Chemie und
Molekulare Strukturbiologie

Wien, im Dezemberr 1999

Much Ado About Nothing...

William Shakespaere ca. 1599

Für meine Eltern

Ich danke an dieser Stelle all jenen, die zum Entstehen dieser Arbeit beigetragen haben.

Prof. Peter Schuster, der mir die ermöglichte meine Dissertation in seiner Arbeitsgruppe durchzuführen. Dr. Peter Stadler dafür, daß er mich dazu ermunterte an diesem Institut meine Diplomarbeit und meine Dissertation unter seiner Betreuung durchzuführen, sowie für seine Anregungen und Ratschläge. Dr. Ivo Hofacker für seine umfangreiche und geduldige Unterstützung, sowohl in wissenschaftlicher Hinsicht als auch in allem Computerfragen, ohne welche die vorliegende Arbeit nie zustande gekommen wäre.

Judith Jakubetz danke ich dafür, daß sie im Hintergrund dafür sorgt daß wir uns so wenig wie möglich um die Bürokratie kümmern müssen und trotzdem alles immer seine Richtigkeit hat. Meinen derzeitigen bzw. ehemaligen Kollegen und Freunden Jan Cupal, Daniela Dorigoni, Martin Fekete, Christoph Flamm, Thomas Griesmacher, Kurt Grünberger, Jörg Hackermüller, Christian Haslinger, Stephan Kopp, Michael Kospach, Bärbel Krakhofer, Stefan Müller, Susanne Rauscher, Alexander Renner, Bärbel Stadler, Roman Stocsits, Andreas Svrcek-Seiler, Caroline Thurner, Günther Weberndorfer, Andreas Wernitznig, Christina Witwer und Stefan Wuchty, danke ich dafür, daß sie am Institut eine Klima geschaffen haben, in dem ich mich in den letzten Jahren so wohl gefühlt habe, daß ich es nicht guten Gewissens Arbeitsklima nennen kann.

Meinem Freund Gerald danke ich vom ganzen Herzen dafür, daß er mich durch mein gesamtes Studium begleitet hat und mir immer mit Rat und Tat zur Seite stand.

Meinen Eltern danke ich für ihre Geduld und ihre Unterstützung während der Jahre meines Studiums.

Zusammenfassung

Knowledge-Based Potentials können verwendet werden um zu entscheiden, ob eine Aminosäuresequenz in eine vorgegebene native Proteinstruktur falten wird. Wir verwenden diesen Ansatz um die Sequenz-Strukturbeziehungen im Proteinraum zu untersuchen. Vorallem untersuchen wir folgende Annahmen, die für eine effiziente Evolution von großer Bedeutung sind: (i) Sequenzen, die in eine vorgegebene native Struktur falten bilden umfangreiche *Neutral Netze*, die den Sequenzraum durchziehen. (ii) Die *Neutralen Netze* zweier nativer Strukturen nähern sich bis auf wenige Punktmutationen. Dies wollen wir mittels Computersimulationen mit zwei völlig unterschiedlichen Potentialfunktionen verifizieren: Manfred Sippl's `PROSA II` Paar Potential und und A. Lapedes' `Neuralem Netz (NN)` Potential. Um die Topologie der neutralen Menge $S(\psi)$ für eine Struktur ψ zu untersuchen, verwende wir die Technik des inversen Faltens zur Entscheidung ob eine Aminosäuresequenz x ein Mitglied von $S(\psi)$ ist, d.h. ob x in die Struktur ψ faltet. Dieses Problem ist weniger anspruchsvoll als die Voraussage einer unbekannten Struktur anhand einer gegebenen Sequenz. Als Maß dafür wie gut die Sequenz x auf die Struktur ψ paßt, verwenden wir dem *z-score*. Formal übersetzen wir das inverse Faltungsproblem in eine ein Optimierungsproblem auf der Menge aller Sequenzen: wir suchen das Optimum von x für den z-score von (x, ψ). Wir finden, daß sich die neutrale Pfade innerhalb der Menge $S(\psi)$ bis beinahe zur Länge der Aminosäuresequenz ausdehnen. Wir schließen daraus, daß neutrale Mengen umfangreiche neutrale Netze bilden, die den gesamten Sequenzraum durchziehen. Experimente zur Untersuchung des kleinstmöglichen Abstands zweier neutralen Mengen innerhalb des Sequenzraums zeigten, daß sich die neutralen Mengen zweier unterschiedlichen Strukturen $S(\psi)$ und $S(\varphi)$ sehr nahe zusammen kommen, wir folgern daraus, daß im Proteinraum *shape space covering* gegeben ist. Wir fanden eine vergleichsweise gute Korrelation zwischen den Ergebnisse aus dem `NN` und dem `PROSA II` Potential. Obwohl im Detail Unterschiede bestehen, fanden sich bei den *adaptive walk*, *neutral walk* und *closest approach walk* Experimenten die gleichen Eigenschaften. Auch die aus diesen Ergebnissen folgende Implikation der Existenz von ausgedehnten neutralen Netzen und *shape space covering* gilt sowohl für `PROSA II` und das `NN` Potential. Folglich sind un-

sere Schlußfolgerungen bezüglich der der Topologie des Sequenzraumes die sich aus den unterschiedlichen Computerexperimenten ergeben unabhängig von den Details des verwendeten Potentials.

Abstract

Knowledge-Based potentials can be used to decide whether an amino acid sequence is likely to fold into a prescribed native protein structure. We use this idea to survey the sequence-structure relations in protein space. In particular, we test the following two propositions which were found to be important for efficient evolution: The sequences folding into a particular native fold form extensive *neutral networks* that percolate through sequence space. The neutral networks of any two native folds approach each other to within a few point mutations. Computer simulations using two very different potential functions, Manfred Sippl's `PROSA` pair potential and Alan Lapedes' neural network based `NN` potential, are used to verify these claims and to test whether the results are independent of the potential used to obtain the results. In order to characterize the topology of neutral sets $S(\psi)$ for the protein structure ψ we use an inverse folding technique to decide whether a given amino acid sequence x is a member of $S(\psi)$, that is, whether x folds into the structure ψ. This problem is less demanding than predicting the unknown structure of a given sequence. As a measure for the quality of fit of sequence x and structure ψ we use the *z-score*. Formally, we translate inverse folding into an optimization problem on the set of all sequences: we are looking for an optima x of the z-score $z(x, \psi)$. We find that neutral paths within the sets $S(\psi)$ extend to almost the length of the amino acid sequence. We therefore conclude that neutral sets form extensive *neutral networks* that percolate the entire sequence space. Our closest approach experiments showed that the neutral sets of two different structures $S(\psi)$ and $S(\varphi)$ come closely together, we therefore conclude that protein space exhibits *shape space covering*. A comparably good correlation between the `NN` and `PROSA II` potential was found. Although some differences appear in detail, the behavior of adaptive walks, neutral walks, and closest approach walks, and consequent implications such as the existence of extensive neutral networks and shape space covering, are common to both the `PROSA II` and the neural network `NN` potentials. Hence our conclusions concerning the topology of sequence space, as defined by the various types of walks, are independent of the details of any one potential.

Inhaltsverzeichnis

1 Introduction **1**

2 Potentials **5**

2.1 Molecular Mechanics Force Fields 5

2.2 Knowledge Based Potentials 9

2.2.1 The Inverse Boltzmann Law 10

2.2.2 Log-Likelihood Ratios 12

2.3 Various Approaches to Knowledge-Based Potentials 13

2.3.1 Atom-Atom Potentials 13

2.3.2 Contact Potentials 14

2.3.3 Profiling Potentials 16

2.3.4 Tropsha's Four-Point Potential 16

2.3.5 Sippl's `PROSA II` 20

2.3.6 Lapedes' Neural Network `NN` Potential 21

3 The z-score **22**

4 Inverse Folding of Proteins **23**

4.1 Inverse Folding 23

4.2 Neutral Sets 24

4.3 Adaptive Walks 25

5 The Protein Structures **28**

5.1 P22 C2 Repressor 28

5.2 Ubiquitin 29

5.3 Calbindin D9K 31

5.4 Thioredoxin 32

5.5 Cystatin 33

5.6 Rat Oncomodulin 34

5.7 Lysozyme 35

5.8 The Janus Proteins 36

6 Adaptive walks **38**

6.1 Adaptive Walks `1ADR` 38

6.2 Adaptive Walks `1UBQ` 39

6.3 Adaptive Walks `4ICB` 42

6.4 Adaptive Walks `1CEW` 47

6.5 Adaptive Walks `1RRO` 50

6.6 Adaptive Walks `2TRXA` 54

6.7 Adaptive Walks `1LYZ` 57

6.8 Summary 61

7 Neutral Networks **64**

8 Closest Approach and Shape Space Covering **70**

9 Janus **74**

10 Conclusion and Outlook **78**

11 List of Figures **82**

12 List of Tables **86**

13 References **87**

1 Introduction

A core problem of modern biosciences is the design of novel proteins with predefined and adjustable functions. As these functions directly reflect a certain structure, i.e., a framework of functional groups in three dimensions, the problem can be expressed differently as how to rationally design a protein structure with only a few side chains (e.g. from the "active site") exactly placed on an optimal scaffold.

Despite the fact that a large body of knowledge on protein folds has accumulated over the past decades, it still remains impossible to calculate native structures from amino acid sequences. Presently many rules are known that are important for the stability of protein chains. In some special cases the structure of proteins has been predicted ahead of experiment. Nevertheless, correct predictions of structures are singular events that depend largely on the knowledge of homologous proteins with known structures. Even an approximate map of protein space will therefore be helpful in protein design since it can be used to direct experimental procedures.

Mapping the sequence-structure relations of RNA, based on secondary structure predictions, has provided a theoretical basis for understanding the dynamics of *in-vitro* evolution (e.g. SELEX) experiments. In particular, the discovery of extended neutral networks in computer simulation provides an explanation why (and how) an evolutionary biotechnology based on functional RNA molecules is feasible at all [52, 37].

Protein space, on the other hand, is still largely *Terra incognita*. Considering the hyper-astronomical number of possible sequences, a detailed mapping of protein space is a hopeless task. On the other hand, the repertoire of stable native folds seems to be highly restricted or even vanishingly small [15, 34]. It makes sense therefore, to ask how the set $S(\psi)$ of all those sequence that fold into the same shape ψ is distributed in sequence space. We call these sequences *neutral*, $S(\psi)$ being the *neutral set* of the fold ψ. The shape or topology of neutral sets has important implications for the evolution of proteins and for *de novo* design. Partial answers to that question were recently obtained by computational studies using

lattice models and so-called *knowledge-based potentials*, as well as by some crucial experimental findings [3, 2, 11, 17].

- Are biological (evolved) structures common or rare in the space of protein sequences?
- Are biologically relevant structures confined to a small connected subspace of protein space or can they be found "all over the place"?
- Is the observed bias in amino acid composition necessary for folding or is it a product of abundance and evolution?
- Is it possible to restrict the set of amino acids to only a few ones, still preserving structure or even function?
- Can we find different protein folds for very similar sequences, and conversely, can we find the same fold for unrelated amino acid sequences?
- Are there *neutral networks* in protein space? That is, is it possible to walk from one end of sequence space to the other, via point mutation steps, and still fold to the same structure at every intermediate?
- Do we have *shape space covering*? That is, is it possible to find almost all relevant folds within a small radius around any randomly chosen reference sequence?

All these questions have been answered with *yes* for the secondary structures of nucleic acids in a series of investigations conducted mostly by Peter Schuster's groups at the Institute for Theoretical Chemistry in Vienna and at the IMB in Jena [22, 23, 24, 53, 60, 36, 30, 31, 61, 37]. In addition, it has been shown that the results are robust with respect to small changes in energy parameters and folding rules [61].

Three approaches have been applied so far to study the topology of neutral sets: a mathematical model of genotype-phenotype mapping based on random graph theory [48], extensive sample statistics [53] using *neutral walks* as a "probe", and exhaustive folding of all sequences with given chain length n [31].

The mathematical model assumes that sequences forming the same structure are distributed randomly using the fraction λ of neutral neighbors as (the only) input parameter. If λ is large enough this model makes two rather surprising predictions [48, 49]:

- The connectivity of $S(\psi)$ changes drastically when λ passes the threshold value:

$$\lambda_{cr}(\alpha) = 1 - \sqrt[\alpha-1]{\frac{1}{\alpha}} \approx 0.146 \tag{1}$$

  where $\alpha = 20$ is the size of the amino acid alphabet. The neutral set $S(\psi)$ consist of a single component that spans the sequence space if $\lambda > \lambda_{cr}$, while it is partitioned into a large number of components below threshold.

- There is *shape space covering*, that is, in a moderate size ball centered at any position in sequence space there is a sequence x that folds into any prescribed secondary structure ψ, see the following section.

In this thesis global properties of the sequence-structure relation of polypeptides are investigated using knowledge based potentials.

Previous computer simulations [3] have shown that knowledge-based potentials can be used in principle to answer questions concerning the sequence-structure relationship of proteins. In fact, knowledge-based potentials are designed to recognize whether a sequence x folds into a native structure ψ. This problem is by far less demanding than predicting the unknown structure of a given sequence because it can be investigated by inverse folding techniques [20, 8].

Recent studies using knowledge based potentials [4, 8, 26, 27, 29, 33, 57, 58] demonstrated that the energy of the native fold (i.e., putative ground state) of a sequence x can be estimated from the distribution of the energy values of x in its conformation space. This allows the construction of an energy scale (z-score) by which conformations of different sequences can be compared. Empirically, native folds have z-scores in a narrow characteristic range. Hence we may assume that x assumes the native fold ψ if the z-score of $z(x, \psi)$ is in the native range.

This thesis extends earlier results in two directions: (i) Repeating earlier computer experiments using a very different potential function, we evaluate the results from one potential in the other potential and vice versa to find common traits of knowledge-based potentials and to verify our previous results. (ii) It answers the main question that was left open in [3], namely whether there is *shape space covering*, that is whether the neutral networks of any two different shapes come close to each other.

In this work we used two very different potential functions, Alan Lapedes's Neutral Network `NN Potential` [12] and Manfred Sippl's `PROSA II` [33, 56, 57, 58, 29], based on quite different encoding of the protein structures two answer questions concerning the sequence-structure relationship of proteins and to investigate whether or not results found in earlier experiments [3] can be verified when reevaluated in another potential.

2 Potentials

The energy of a macromolecular system is a function of the conformational variables (e.g. Cartesian coordinates) plus its interaction energy with the surrounding solvent. The derivation of the energy from the conformational variables gives the force field of the molecule. The term potential in this context is a synonym to energy function. Generally we assume that a protein sequences $s = (s_1, \ldots, s_n)$ of n amino acids

$$s_i \in \{\mathbf{A}, \mathbf{C}, \mathbf{I}, \mathbf{L}, \mathbf{M}, \mathbf{F}, \mathbf{W}, \mathbf{Y}, \mathbf{V}, \mathbf{R}, \mathbf{N}, \mathbf{D}, \mathbf{E}, \mathbf{Q}, \mathbf{G}, \mathbf{H}, \mathbf{K}, \mathbf{P}, \mathbf{S}, \mathbf{T}\}$$

is related with its structure ψ as represented by the coordinates $x_s = (x_1, \ldots, x_n)$ via the potential function $V(s, \psi)$:

$$x_s = \text{argmin}_x V(s, \psi)$$

The design of molecular force fields allows at least two different approaches:

On the one hand semi-empirical approaches consider macromolecular systems as a summation of the forces observed for monomers. The force fields are obtained from quantum mechanical calculations, and data from thermodynamic or spectroscopic measurements on small molecules.

On the other hand knowledge-based potentials are based on the assumption that force fields of macromolecules are of immense complexity and the only reliable source of information are macromolecular molecules themselves. So empirical or knowledge-based potentials try to extract information from databases of macromolecular structures.

2.1 Molecular Mechanics Force Fields

The need to describe molecular structures and properties in a practical manner led to the development of the "mechanical" molecular model [46, 65]. Although highly accurate, quantum chemical calculations necessitate a computational effort so immense, that solving the Schrödinger equation for macromolecular systems

is impossible for the time being. Hence, to calculate atomic structures a classical approach based on the following assumptions had to be chosen :

- According to the Born-Oppenheimer approximation of the Schrödinger equation, nuclei and electrons are aggregated in atom-like particles.
- These particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Atoms are considered as balls, bonds represented by springs allowing the use of classical potential functions.
- Interactions must be preassigned to specific sets of atoms.
- The spatial distribution and energies of particles are determined by interactions

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanic energies are no absolute quantities. Only the energy difference between two conformations of the same molecule are meaningful. A simple molecular mechanic energy equation is given by:

$$E_{tot} = E_{stretch} + E_{bend} + E_{tors} + E_{non-bonding}$$

These terms together with the parameters required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model. The constants (force constants, equilibrium lengths) can be either measured by spectroscopy or calculated by quantum mechanical means.

These terms together with the parameters required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model. The constants (force constants, equilibrium lengths) can be either measured by spectroscopy or calculated by quantum mechanical means.

The energy terms in detail are:

Stretching Energy:

Occurs whenever a bond is deformed (stretched or compressed), and is described by an equation based on Hooke's law for springs.

$$E_{stretch} = \sum k_b(r - r_0)^2$$

whereby k_b is the force constant, r is the actual bond length and r_0 the equilibrium length. This parabolic approximation fails as the bond is stretched toward the point of dissociation.

Bending Energy:

Energy increases if the equilibrium bond angles are bent. Again the approximation is harmonic and uses Hooke's law.

$$E_{bend} = \sum k_\theta(\theta - \theta_0)^2$$

k_θ controls the stiffness of the angel, θ is the actual bond angle, θ_0 the equilibrium angle. The force constants have to be estimated for each triple of atoms (e.g. C-C-C, C-C-O, C-C-H)

Torsion Energy:

Intra-molecular rotations (around torsions or dihedrals) require energy as well:

$$E_{torsion} = \sum A[(1 + \cos(n\tau - \phi))]$$

The parameter A controls the amplitude of this periodic function, n the periodicity, and ϕ shifts the entire curve along the rotation angle axis

τ. Again the parameters for all combinations of four atoms have to be determined (e.g. C-C-C-C, C-O-C-C, H-C-C-N).

Non-bonding Energy:

The different implementation of force field differ mainly in the definition of this term. Mostly present are Van der Waals and electrostatic terms.

$$E = \underbrace{\sum_i \sum_j -(\frac{A_{ij}}{r_{ij}^6} + \frac{-Bij}{r_{ij}^{12}})}_{\text{Van der Waals}} + \underbrace{\sum_i \sum_j \frac{q_i q_j}{r_{ij}}}_{\text{Coulomb}}$$

The Van der Waals term accounts for the attraction and the Coulomb term for electrostatic interaction. Repulsion occurs, when the distance between two atoms becomes less than the sum of their radii. The shown approximation for the van der Waals energy is of the Lennard-Jones potential type. It is used this way for instance in in the AMBER force field [65] as can be seen in equation(2). The last term accounting for H-bonds is modeled by a 6-12 potential as well.

$$\begin{aligned} E_{\text{total}} = & \sum_{\text{bonds}} K_r(r - r_{eq})^2 \\ & + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \\ & + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ & + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \\ & + \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] \end{aligned} \tag{2}$$

2.2 Knowledge Based Potentials

In contrast to the analytic approach of mechanical force fields, knowledge based potentials describe the energy needed for a certain contact to occur by a likelihood [4, 8, 26, 27, 29, 33, 57, 58]. This likelihood of finding a particular contact is extracted from a database of known structures. Computer scientists call this procedure *data-mining*. The increase of information is measured by the log-likelihood ratio of the Bayesian events [5]. This ratio is the relation of prior expected events and the observed occurrence. Therefore the log-likelihood is a kind of measure for the "surprise" provided by the database.

A physical interpretation of the probability function comes from statistical mechanics: Based on the assumption that the protein is in its energetic minimum, low energy elements must occur more frequently than others in $3d$-structures of globular proteins. This dependence of occurrence on energy resembles a Boltzmann statistic [12, 56]:

$$f_{occ.} \sim \exp\left(-E/RT\right)$$

Here T is the conformational temperature and R is the gas constant. This similarity reveals, that if in principle the frequency of occurrence can be estimated, it is possible to gain access to the putative energy of a certain fold $\psi(S)$. This interpretation of knowledge based potentials was introduced by Manfred Sippl and is the basis for most of the contemporary potentials of mean force.

Recently Dill and Thomas stated severe critique on this approach of statistical potentials [62]. They intended to test how "extracted" energies correspond with "true" energies by mimicking the extraction process on ideal lattice models and comparing the observed with the accurate energy of **HP** interactions. Their major points of criticism for this model are that proteins are not seen as chains (rather as gas composition) and the temperature applied to the Boltzmann device is meaningless. Further they try to show that the energies for a certain fold depend solely on clustering of polarity. These findings were put into theoretic framework recently by Neumaier's *"Non-uniqueness Theorem"* [45]. It shows, that empirical potentials obtained by extraction of equilibrium geometries can *never* reveal *true* energies. In particular, empirical potentials derived solely from databases of

equilibrium data will never be useful for dynamical studies.

2.2.1 Statistical Thermodynamics of Proteins or the Inverse Boltzmann Law

The so called "*folding postulate*" states, that "*In equilibrium the native state of a protein-solvent system corresponds to the global minimum of free energy*". This was demonstrated in the pioneer study performed by Anfinsen [1] in 1973. He was able to show, that by reducing and re-oxidating disulfide bonds in ribonuclease no loss of function occurs, i.e. that folding is a reversible process.

In the following derivation that follows essentially Sippl [56], the peptide chains will be presented by C^{α} atoms to make the model easier, by no loss of generality. According to Bolzmann's law the probability $f(x)$ of finding a physical system in a particular state x in equilibrium is give by

$$f(x) = \frac{1}{Z} \exp\left[-\frac{E(x)}{kT}\right]$$

Where k is the Boltzmann's constant, T the absolute temperature in Kelvin (Reference temperature) and Z is the partition function defined as

$$Z = \int \cdots \int \exp\left[-\frac{E(x)}{kT}\right] dx$$

For discrete systems the integral may be replaced by the sum.

$$Z = \sum_{x=1}^{n} \exp\left[-\frac{E(x)}{kT}\right]$$

If the energies of all states x were known, the probability density could be computed. On the other hand it is possible to obtain the energy if the density of states can be measured [56].

$$E(x) = -kT \ln\left[f(x)\right] - kT \ln Z \tag{3}$$

From equation (3) it is possible to calculate the energy of a particular distribution but it is impossible to get the Boltzmann sum Z, so an additive constant remains unknown. If the probabilities of a distribution are extracted from a database,

the potential of mean force of interaction can be obtained. If $E(x)$ denotes the reference state of the system (averaged energy), the net potential for a given interaction γ can be computed by:

$$\Delta E_\gamma(x) = E_\gamma(x) - E(x)$$

or:

$$\Delta E_\gamma(x) = -kT \ln \left[\frac{f_\gamma(x)}{f(x)} \right] - kT \ln \frac{Z_\gamma}{Z}$$

and since Z_γ and Z do not depend on the state x, it is legitime to assume $Z_\gamma \simeq Z$, and therefore $-kT \ln \frac{Z_\gamma}{Z} \sim 0$. T is tied to the temperature of the NMR or X-ray measurement of the data.

$$\Delta E_\gamma(x) = -kT \ln \left[\frac{f_\gamma(x)}{f(x)} \right]$$

Due to the restriction of a limited number of observations it must be distinguished between the probability densities $f(x)$ or $f_\gamma(x)$ and the information obtained from the database $g(x)$ respectively $g_\gamma(x)$. It is reasonable however to approximate the reference state probability $f(x)$ with $g(x)$ since the overall number of interactions in the database is big enough (magnitude of 10.000). On the other hand the number of observations can be low for particular contacts, especially when considering higher order interactions. Therefore database size is crucial for the approximation of $f_\gamma(x) \approx g_\gamma(x)$.

So without knowledge of any specific interaction we have to assume $f(x) \approx f_\gamma(x)$ and expect $\Delta E_\gamma(x) \equiv 0$. Each information quantum derived form the database increases $f_\gamma(x)$, and the net contribution is twofold: (1) The relative energy of all states $\Delta E_\gamma(x)$ is increased and (2) the energy of a particular state $\Delta E_\gamma(t)$ is lowered. This means that if $f_\gamma(x) < 1$ the contribution to the overall energy becomes negative. When parameters for all configurations γ are extracted, a summation over all contributions yields the energy of sequence S for structure ψ:

$$E(S, \psi) = \sum_\gamma E_\gamma(x)$$

Over the past years many different approaches to potentials of mean force have been made. The various potential functions are distinct in the definition as well as

in the order of interaction. Therefore different "resolutions" are used to define the energy functions. The spectrum reaches from an atomic resolution mode (Sippl) to simplified **HP**-patterns (Crippen), and a lot in between.

Munson *et al.* [44] were able to show that increasing the order of interaction improves the statistical significance of the terms. Starting with a highly significant one body term, that counts for the exposures of the residue, continuing to a pair potential term, that contributes for amino acid preferences (e.g. hydrophobic-hydrophobic interactions) independent of the burial status, one can clearly identify that multi-body interactions participate to a major extent the overall potential function.

2.2.2 Log-Likelihood Ratios

Empirical Potentials can also be considered from a statistical instead of a statistical mechanical view [10, 29]. The statistical approach involving log-likelihood potentials for the construction of of potential functions was introduced by Bryant and Lawrence [10].

For example in Sippl's [56] `PROSA II`, the probability for amino acids pairs ab to be separated by a (binned) distance r, is approximated by counting the frequency in a database of know protein structures. This probability can then be represented as the conditional probability $P(r|ab)$ of finding the distance bin r for all 20×20 possible amino acid pairs. The approximation for statistical mechanical free-energies is then represented by the following:

$$\log\left(\frac{P(r \mid ab)}{P(r)}\right) = \log\left(\frac{P(r, ab)}{P(r)P(ab)}\right)$$

The interpretation from a statistical point of view, quantifies the relation between $P(r)$ and $P(ab)$ by comparing the joint probability $P(r, ab)$ to the probability obtained under the assumption of independence between the distance r and the pair ab. The log-likelihood quantities obtained from a *'training set'* of proteins can then be used to evaluate the compatibility of any specific amino acid sequence

with any given structure. Different types of potential function are based on different log-likelihood expressions. While Wilmanns and Eisenberg [66] express the log-likelihood ratio for their potential function as $\log\left(\frac{P(ab|r)}{P(r)}\right)$, Sippl's ratio [56] is given as $\log\left(\frac{P(r|ab)}{ab}\right)$. However when related under Bayes theorem:

$$\log\left(\frac{P(r \mid ab)}{P(r)}\right) = \log\left(\frac{P(ab \mid r)}{P(ab)}\right)$$

these are identical expressions [29]. Still different log-likelihood expressions are used by Bryant and Lawrence [10] and Skolnik et. al. [26]. In Alan Lapedes `NN` potential the probabilities of the various pairs ab to have inter-residue distances within certain distance bins are computed by similar frequency counting, but the log-likelihood ratio $\log\left(\frac{P(ab|r)}{P(a)P(b)}\right)$ differs from the above, due to the fact that the probabilities are compared to the probability assuming the sequence was randomly permuted and re-threaded.

2.3 Various Approaches to Knowledge-Based Potentials

2.3.1 Atom-Atom Potentials

The reversible energy required to bring two particles close to each other at constant volume is given by the potential of mean force or Helmholtz free energy of the system. It is related to the radial distribution function $g(r)$ through

$$w(r) = -kT \ln[g(r)]$$

and can give insights to protein folding and the role of specific interaction in native structures (e.g. H-bonds). The distribution function for arbitrary sets of atom-atom interactions occurring in proteins can either be obtained by diffraction experiments, or they are extracted from a database of structures. The two functions turn out to be equal, if the distance distributions are similar. The knowledge based distribution function is accessed by the determination of

$$\rho_{ab}(r) = \sum_{ab} \delta(r - r_{ij})$$

as the sum over all distinct pairs ab within the radius r in a protein library. The observed density is compared with a bulk of non interacting particles to finally obtain the distribution function:

$$g_{ab}(r) = \frac{\rho_{ab}(r)}{\rho}$$

The potentials using these distribution functions are perfectly suited for a detailed analysis of spatial distributions of atom contacts along a protein chain [59]. To make use of an atom-atom based potential, one has to know the Cartesian coordinates for *all* residues in a poly peptide chain. Therefore this approach is of no use to solve the inverse folding problem, as targeted by our group.

2.3.2 Contact Potentials

Contact potentials can be understood as subgroup of knowledge based potential. This kind of mean energy function measures the overall energy of a system, as the sum of *nearest neighbor* contacts. Some of the early works considered the frequency with which pairs of amino acids appeared within a certain "contact" distance of each other and used a quasi-chemical approximation to relate this frequency to an approximate free-energy of interaction of a "gas" of residue pairs [43].

One of the most prominent examples is Crippen's Simplified Potential. To obtain a simplified representation of heteropolymers Ken A. Dill introduced the concept of lattice polymers [13]. When used to model proteins, each amino acids occupies one position on the grid of the lattice. Conformations of lattice polymers are represented by *self-avoiding walks*, short SAWs. Hence this method greatly reduces the conformational space of the optimization problem. On a lattice bond lengths are, of course, always constant, furthermore potentials for lattice proteins usually neglect bond angles and dihedrals. Instead they focus on non-bonding interactions of topological neighbors.

In Crippen's potential the energy for the pair interaction has the form

$$E(s, \mathbf{x}) = \sum_{i,j} \Psi[s(i), s(j); |i - j|; d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)]$$

The individual interaction terms Ψ depend on the type $s(i)$ and $s(j)$ of residues,

on their separation $|i-j|$ along the chain and on the euclidian distance $d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)$ of the lattice points. The potential function

$$\Psi[s(i), s(j); |i-j|; d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)] = U[s(i), s(j); |i-j|]g(d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j))$$

is normalized such that the contribution of the nearest neighbor reduces to $U[s(i), s(j); |i-j|]$.

Crippen extracted a contact matrix of the form:

$$U[s(i), s(j); |i-j|] =$$
$$\begin{cases} -0.008 & \text{if } |i-j| = 3 \\ 0.004 & \text{if } |i-j| = 4 \\ 0.021 & \text{if } |i-j| = 5, 6, 7 \\ \begin{pmatrix} -0.012 & -0.074 & -0.054 & 0.123 \\ -0.074 & 0.123 & -0.317 & 0.156 \\ -0.054 & -0.317 & -0.263 & -0.010 \\ 0.123 & 0.156 & -0.010 & -0.004 \end{pmatrix} & \text{if } |i-j| \geq 8 \end{cases}$$

from a structural database where the matrix entries correspond to the four amino acids classes:

1 = {**G Y H S R N E**}
2 = {**A V**}
3 = {**L I C M F**}
4 = {**P W T K D Q**}

A further simplification of the potential can be obtained by restricting the amino acid alphabet to just two classes: **H** for hydrophobic amino acids and **P** for polar residues. For a review of **HP** based potentials see [14, 19].

Crippen recently used the described potential in kinetic simulations and calculations of denaturation curves [16]. These computer experiments showed that folding kinetics largely depend on the coding scheme and that the results obtained by using the Crippen alphabet differ strongly from calculations for spin-glass encoded SAWs [27, 28].

2.3.3 Profiling Potentials

Eisenberg and coworkers decided to "translate" the $3d$-structures to a $1d$-string, using three parameters:

1. The total side-chain area being covered by any other protein atoms
2. The fraction of side-chain area being covered by polar atoms or water molecules
3. The local secondary structure

The environment strings were extracted from a database of known structures. The resulting environment classes discriminate buried and exposed residues, and further subdivisions yield 18 distinct classes for the 20 amino acids. The optimization problem was to find the most favorable alignment of a protein sequence to the environment string, whereby classical alignment techniques came to use. The resulting threading procedure has been successfully employed to identify sequence-structure pairs.

2.3.4 Tropsha's Four-Point Potential

Avoiding the arbitrariness of a binned distance, A. Tropsha [68, 55, 69] introduced an approach from computational geometry to knowledge based potentials. He suggested to represent the protein structure as a set of points in $3d$, for simplification only C^{α} atoms were chosen as model for the backbone. This set of points is tessellated using the *Delauney triangulation*. The result of this geometric procedure is a partitioning of the space included by the set into irregular tetrahedra with the points as vertices. The quadruple of amino acids represented by these points are considered to be nearest neighbors. The beauty of this method is that it is parameter free, the list of tetrahedra is non-ambiguous.

If one counts the occurrence of all possible neighborhood combinations of the amino acids in a structural dataset, a log-likelihood function can be constructed

easily. This function can then be used to test if a given sequence yields favorable contacts when threaded to a certain structure — in one word *inverse folding*.

The common meaning of *"tessellation"* is to arrange squares in a mosaic pattern. The term derives from the Greek word *"tesseres"* which means four. Generally *tessellating* can be understood as arranging regular polyhedra congruently (all angles and sides are equal) in a plane with edges attached to each other. Only three regular polygons tessellate in the Euclidean plane: triangles, squares and hexagons (see figure 1). By extension, space or hyper space may also be tessellated.

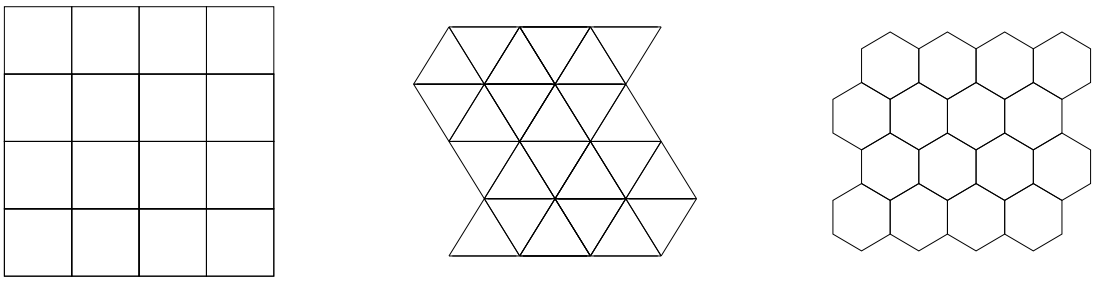

FIGURE 1: Tessellations in two dimensions.

The Delauney triangulation *tessellates* a set of points in $\mathbb{R}^3$ in the sense of filling space with tetrahedra. The Delauney triangulation is computed via its dual, the Voronoi diagram.

Given a set S of n distinct points in $\mathbb{R}^d$, a Voronoi diagram is the partition of $\mathbb{R}^d$ into n polyhedral regions $\mathrm{vo}(p), (p \in S)$. Each region $\mathrm{vo}(p)$, called the Voronoi cell of p, is defined as the set of points in $\mathbb{R}^d$ which are closer to p than to any other points in S, or more precisely,

$$\mathrm{vo}(p) = \{x \in \mathbb{R}^d | \mathrm{dist}(x,p) \leq \mathrm{dist}(x,q) \forall q \in (S-p)\}$$

where dist is the Euclidean distance function. The set of all Voronoi polyeders forms a cell complex. The vertices of this complex are called the *Voronoi vertices*, and the extreme rays (i.e. unbounded edges) are the *Voronoi rays*.

For each point $v \in \mathbb{R}^d$, the *nearest neighbor* set $\mathrm{nb}(S,v)$ of v in S is the set of points $p \in S - v$ which are closest to v in Euclidean distance. In order to compute the Voronoi diagram, the following construction is very important. For

each point p in S, consider the hyper-plane tangent to the paraboloid in $\mathbb{R}^{d+1}$: $x_{d+1} = x_1^2 + \cdots + x_d^2$. This hyper-plane is represented by $h(p)$:

$$\sum_{j=1}^{d} p_j^2 - \sum_{j=1}^{d} 2p_j x_j + x_{d+1} = 0$$

By replacing the equality with inequality $\geq$ above for each point p, we obtain the system of n inequalities, which we denote by $b - Ax \geq 0$. The polyhedron P in $\mathbb{R}^{d+1}$ of all solutions x to the system of inequalities is a lifting of the Voronoi diagram to one higher dimensional space. In other words, by projecting the polyhedron P onto the original $\mathbb{R}^d$ space, we obtain the Voronoi diagram in the sense that the projection of each facet of P is associated with exactly the Voronoi cell $\mathrm{vo}(p)$. The vertices and the extreme rays of P project exactly to the Voronoi vertices and the rays, respectively.

Let S be a set of n points in $\mathbb{R}^d$. The convex hull $\mathrm{conv}(nb(S, v))$ of the nearest neighbor set of a Voronoi vertex v is called the Delauney cell of v. The Delauney complex (or triangulation) of S is a partition of the convex hull $\mathrm{conv}(S)$ into the Delauney cells of Voronoi vertices.

The Delauney complex is not in general a triangulation but becomes a triangulation when the input points are non-degenerate, i.e. no $d+2$ points are cospherical or equivalently there is no point whose nearest neighbor set has more than $d+1$ elements. The Delauney complex is dual to the Voronoi diagram in the sense that there is a natural bijection between the two complexes which reverses the face inclusions.

There is a direct way to represent the Delaunay complex, just like the Voronoi diagram. In fact, it uses the same paraboloid in $\mathbb{R}^{d+1} : x_{d+1} = x_1^2 + \cdots + x_d^2$. Let $f(x) = x_1^2 + \cdots + x_d^2$, and let $\tilde{p} = (p; f(x)) \in \mathbb{R}^{d+1}$ for $p \in S$. Then the so-called lower hull of the lifted points represents the Delauney complex. More precisely, let

$$P = \mathrm{conv}(\tilde{S}) + \mathrm{noneg}(e^{d+1})$$

where e^{d+1} is the unit vector in $\mathbb{R}^{d+1}$ whose last component is 1. Thus P is the

unbounded convex polyhedron consisting of conv($\tilde{S}$) and any nonnegative shifts by the "upper" direction r. The nontrivial claim is that the the boundary complex of P projects to the Delauney complex: any facet of P which is not parallel to the vertical direction r is a Delauney cell once its last coordinate is ignored, and any Delauney cell is represented this way.

Considering a set of points in $\mathbb{R}^3$ the Delauney triangulation describes an algorithm to decompose the convex hull of these points into tetrahedra.

As previously described, the first step in generating the tessellation built from the irregular tetrahedron is finding the convex hull, which is the smallest convex set of points containing the entire set. The hull is represented by a set of facets and a set of adjacency lists giving the neighbors and vertices for each facet. In $\mathbb{R}^3$ facets are triangles and ridges are edges. The Delauney triangulation in $\mathbb{R}^d$ is calculated from a convex hull in $\mathbb{R}^{d+1}$ by lifting the points to a paraboloid by adding the sum of the squares of the coordinates and computing their convex hull, the set of ridges of the lower convex hull is the Delauney triangulation of the original set.

The `qhull` algorithm [9] is a variation of the randomized incremental algorithm, employing a constructed additional point at the hull to decide which facet belongs to it. The point is outside the facet if it is above the set and in the `qhull` variation of the original version, the point is not created randomly, but at the furthest distance from the outside set. This method is used in the program `qhull` which is publically available via the Internet [1)]. It has been shown empirically [9] that this algorithm is especially efficient and well suited for a $3d$ set of points.

This algorithm of triangulation can be applied to any set of points in space, always objectively describing neighborhood. Representing amino acids of a polypeptide chain by an atom (e.g. C^α or C^β) leads to a regular set of points in $3d$ space, that can be tessellated applying the rules described above. The Voronoi polyhedron is now the region around an atom, each side describes a contact to a neighbor. The underlying Delauney simplices are irregular tetrahedra with an amino acids at each corner. This diagram can be employed to describe contacts

[1)] URL: `http://www.geom.umn.edu/software/download/qhull.html`

of amino acids objectively in $3d$ space.

Tropsha's potential was extended by Günther Weberndorfer [63, 64] for his Master Thesis at our Institute. He developed the Vienna Tessellation Potential. Like the `PROSA II` potential and in contrast to Tropsha's original version, it uses the C_β atoms to represent the amino acid residue (interpolating a virtual C_β atom for Glycine) and furthermore, a special term for surface contacts was introduced. These extensions much improved the quality of this potential function supplying us with another tool to use for the exploration of protein space topology. Consequently a number of computer experiments performed with the `PROSA II` and the `NN` potential were reproduced with the Vienna Tessellation Potential.

2.3.5 Sippl's `PROSA II`

`PROSA II` is a true pair potential with an additional surface term. It was designed to evaluate experimentally determined structures of globular proteins , to identify incorrectly folded proteins (or sections of proteins), and as an independent method for evaluating theoretical models of protein structures [12, 33, 56, 57, 58]. It is of the form

$$W_\gamma(x,\psi) = \sum_{i<j} W_\gamma[x_i, x_j, |i-j|; \mathbf{d}_{ij}^\gamma] + \sum_i V_\gamma[x_i; \chi(i)]\,. \tag{4}$$

The additive pair-contributions $W_\gamma[a,b,k;r]$ depend on the type $\gamma = C^\alpha$ or C^β of the backbone atom, on the amino acids $a = x_i$ and $b = x_j$ at the positions i and j of the sequence x, on their separation $k = |j-i|$ along the chain, and on the Euclidean distance $r = \mathbf{d}_{ij}^\gamma$ between the backbone atoms. The surface term $V_\gamma[a;\chi]$ depends on the type γ of the backbone atom, the amino acid $a = x_i$ at sequence position i and the number χ of protein atoms within a sphere centered at the backbone atom of amino acid x_i. The surface term is motivated by the observation that the solvent exposure of an amino acid can be used to model the energetic features of solvent-protein interactions [7, 8, 40]. The parameter χ serves as a (crude) quantitative measure for the surface-exposure of residue a. The values of the `PROSA II` potential listed throughout the paper refer to the C^β backbone. The `PROSA II` potential can be used to calculate the *z-score* for a

given sequence/structure pair see section 3.

2.3.6 Lapedes' Neural Network NN Potential

The NN Potential includes multi-body interactions [29]. The parameterization is based on the notion of a "local neighborhood" of each residue. The database of crystal structures contains atomic information on the location of atoms of residues, as well as the backbone chain to which each residue is connected. Each residue is attached to the backbone of a protein at the C_α position. There are two "special directions" defined by the relative positions along the backbone of the two neighboring atoms, N and C, to the central C_α atom. These two vectors define a plane, to which the normal vector may be erected, thereby providing an invariant three dimensional coordinate system at each C_α. Any residue within an interaction radius of e.g. 8Å to any C_α atom can be labeled with invariant x, y, z coordinates. To solve the problem of how to usefully order the list of coordinates of the spatial neighbors the sphere surrounding each C_α is divided into a small number of finite spatial bins and the identity/occupancy of amino acids of spatially neighboring residues is noted for each bin. The bins are constructed by using the octants of the sphere, which is further divided into two radial shells, one from 0Å to 6Å and the second from 6Å to 8Å. Neighbors along the chain are included because they contain information on local secondary structure, which is ultimately weighted by the neural network in an automatic fashion. An integer valued vector which is essentially the residue composition of each spatial bin therefore serves to invariantly represent the geometrical location and identity of spatial neighbors within each sphere. The contents of local neighborhoods from a database of sequences with little homology is used to train a neural network using backpropagation and the relative-entropy error function to distinguish native from non-native configurations. Alan Lapedes *et al.* [29] developed a potential with multi-body interactions, parameterized in "local neighborhoods" for each residue. He generalized other threading approaches, and ended up in a statistical interpretation. To employ a neural net for finding a log-likelihood ratio containing higher order terms of interaction, it is necessary to find a suitable representation of the available structural information. To tackle this problem an internal coordi-

nate system is defined, setting the C^{α}-atom to the center, and constructing two vectors pointing to the neighboring chain atoms: C and N. This plane has been shown to have an almost constant angle, and a third dimension is spanned by the cross product of $\overrightarrow{C^{\alpha}N} \times \overrightarrow{C^{\alpha}C}$. Further a binned sphere is constructed around the center (C^{α}-atom) of the coordinate system, representing a "neighborhood shell" of residues. To order this shell to spatial residues, the sphere is split into a predefined number of finite, binned sub-shells.

The chain neighbors, carrying information necessary for secondary structure, can be included as well. The M bins are filled with integers mimicking the 20 amino acids, describing the surrounding of a particular C^{α} atom. The neural net is trained on the pdb-select database, and parameters as number of sub-bins, bin size, or bin resolution were varied. Approaches using C^{β} as a core atom showed better results in threading experiments.

3 The z-score

The quality of knowledge-based potentials can be assessed by the socalled z-score, to test how well the potentials differentiate the native fold of a protein from an ensemble of misfolded structures [67]. The z-score is calculated by:

$$z(x, \psi) = \frac{W(x, \psi) - \overline{W}(x)}{\sigma_W(x)} \,. \tag{5}$$

where $W(x, \psi)$ is the energy of the native structure of a protein, $\overline{W}(x)$ is the average energy of sequence x in all conformations (misfolds) in a database and $\sigma_W(x)$ is the standard deviation of the corresponding distribution. Normalization of energies is necessary since the relative ground state energies of different sequences are not available. The z-score introduces a proper normalization, where the range of values of native folds is known [12]. Conversely, this z-score can be used as an approach to *inverse folding*: Given a fixed conformation Ψ one could search for sequences x_i that give z-scores $z(x, \Psi)$ close to the z-score of the native sequence x.

4 Inverse Folding of Proteins

4.1 Inverse Folding

The native structure ψ of a given amino acid sequence x corresponds to the minimum of its free energy, $W(x, \psi)$. If this energy function W were known the native fold could in principle be predicted from the amino acid sequence by energy minimization in conformation space. Although the energy function is complex and the computational problems are formidable, this is in principle a straightforward recipe. It has indeed been used successfully to investigate the sequence-structure relation for RNA molecules [54].

Inverse folding is, *not* just minimization of the energy function in sequence space for a given conformation. This would be the case only if the energy function were normalized such that the native state (ground state) of *each* sequence is equal to 0. This, of course, amounts to solving the protein folding problem for each possible sequence first. As a consequence exploring sequence space seems to be even more demanding than the folding problem.

However, in our previous work [3] we have established that the optimization problem can be solved using an inverse folding approach procedure known as *adaptive walk*, in which a randomly chosen sequence position is mutated and the mutation accepted if the z-score improves.

The existence of *Neutral Paths* and *Neutral Networks* in Protein Space similar to the RNA case was studied with both the `PROSA II` and `NN` potentials using a neutral walk algorithm .

The size of protein space makes it virtually impossible to check directly whether the neutral sets $S(\psi)$ form extensive connected networks, or whether they consist of a large number of disconnected isolated clusters. However, the existence of very long neutral paths suggest that extensive *neutral networks* of sequences folding into the same structure percolate the entire sequence space [48]. The existence of extensive neutral networks meets a claim raised by Maynard-Smith [42] for protein spaces that are suitable for efficient evolution. The evolutionary implications

of neutral networks are explored in detail in [37]. Empirical evidence for a large degree of *functional* neutrality in protein space was presented by Wain-Hobson and co-workers [41]. In previous studies we have introduced *neutral paths* as a tool to measure the connectedness of neutral sets [53, 3]. The usefulness of this approach is also demonstrated in [25].

As a measure for the quality of fit of sequence x and structure ψ we use the *z-score* defined in equation (5) [12]. For the `PROSA II` potential we use the same database as in [12]. Formally, we can now translate inverse folding into an optimization problem on the set of all sequences: we are looking for an optima x of the z-score $z(x, \psi)$.

4.2 Neutral Sets

Sequences belonging to the same fold ψ form a subset $S(\psi)$ of sequence space. These sequences are called *neutral*, $S(\psi)$ being the *neutral set* of fold ψ. The shape or topology of neutral sets has important implications for the evolution of proteins and for *de novo* design [3].

In order to characterize the topology of neutral sets $S(\psi)$ we need a technique for deciding whether a given sequence x is a member of $S(\psi)$, that is, whether x folds into the structure ψ. This problem is less demanding than predicting the unknown structure of a given sequence. It can be investigated by inverse folding techniques [20, 8].

Neutral paths provide a convenient tool to study the properties of $S(\psi)$. A neutral path starting at a sequence x_0 folding into a structure ψ consists of sequences $x_1, x_2, \ldots$ such that

(i) the sequences x_i is obtained by a single point mutation from x_{i-1} for all $i > 0$,

(ii) all sequences x_i fold into ψ, and

(iii) the Hamming distance $d_H(x_0, x_i) = i$, i.e., each mutation increases the distance from the starting point [53].

Since we have not solved the folding problem, we have to resort to a slightly weaker notion of neutrality, we accept a sequence x_i as folding into the prescribed structure ψ if its z-score is similar or better than the wild-type score z^*. A neutral path ends after $\mathcal{L} \leq n$ steps when no mutant of $x_\mathcal{L}$ can be found that has Hamming distance $\mathcal{L}+1$ from the starting point and folds into ψ.

The usefulness of this approach is demonstrated in [25]. The data we have accumulated shows that there are indeed extensive neutral paths, and consequently also neutral networks in protein space. Some data is listed in Table 13. We found that neutral paths within the sets $S(\psi)$ extend to almost the length of the amino acid sequence. We conclude that neutral sets therefore form extensive *neutral networks* that percolate the entire sequence space. The results of our neutral walk simulations are discussed in detail in Chapter 7.

4.3 Adaptive Walks

From the computational point of view, an adaptive walk is the simplest heuristic to find the optima x of the z-score $z(x, \psi)$. It is sufficient to repeatedly try random mutations that are accepted if and only if the z-score improves, see Figure 3. In this study we use only point-mutations. The frequency of amino acids in randomly generated sequences, are the natural frequencies of the amino acids in known proteins.

While the procedure would eventually terminate in a local optimum, in practice we terminate the algorithm at a predefined threshold score z^* . In all cases we choose z^* identical or 2 to 5 z-score units better than the z-score of the wildtype sequence/structure pair. In the case of the `PROSA II` potentials we require that both the C^α and the C^β z-scores improve with each step of the adaptive walk.

Adaptive walks already yield some insight into the structure of protein space. The length ℓ of adaptive walks, that is, the number of accepted steps until z^* is reached, gives information about the ruggedness of the energy landscape [38]. Longer walks imply smoother surfaces with few local optima.

In Table 1 it is shown that adaptive walks with the `NN` potential are consistently

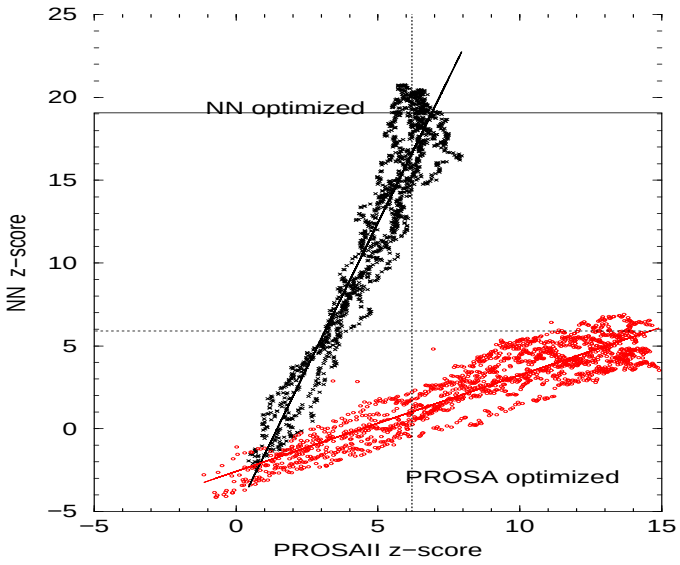

FIGURE 2: Comparism of adaptive walks with the `NN` potential and the `PROSA II` potential.

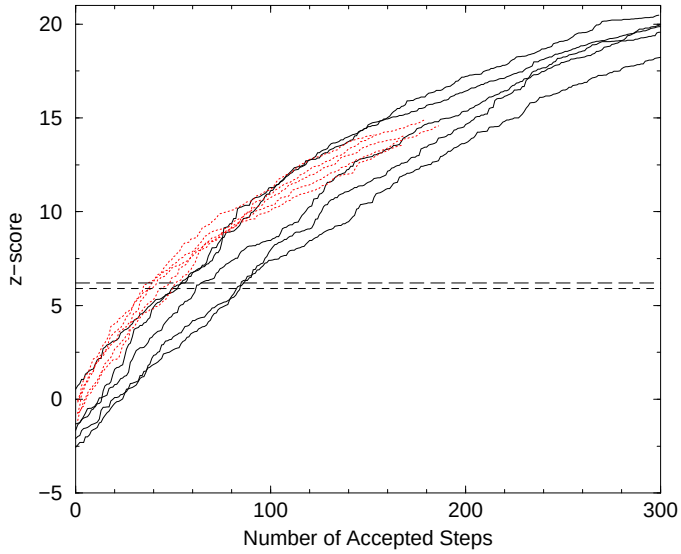

FIGURE 3: Adaptive walks with the `NN` potential (solid lines) and the `PROSA II` potential (dotted lines).

longer than walks on the `PROSA II` surface. Thus the `NN` potential surface contains fewer sequences with wildtype like z-scores, that is, the `NN` potential yields in general a smaller neutral set than the `PROSA II` potential (see Figure 3). The results of all our adaptive walks experiments will be discussed in detail in section 5.8.

TABLE 1: Average length ℓ of Adaptive Walks to reach wildtype z-score.

	PROSA II		NN	
Protein	ℓ	ℓ/n	ℓ	ℓ/n
1cbn	18.7	0.406	−−	−−
1ubq	61.9	0.814	75.4	0.992
1adr	31.7	0.417	−−	−−
4icb	60.3	0.793	75.4	0.992
2trxa	71.7	0.664	112.0	1.037
1rro	79.1	0.732	125.6	1.163
1cew	44.1	0.408	76.0	0.703
1lyz	58.2	0.451	115.2	0.893

The length of the walk ℓ is averaged over 5 runs. ℓ/n is the average length walk normalized by the number of amino acids n in each sequence.

In order to compare the predictions from both potentials we have taken adaptive walks computed with one potential and re-evaluated the sequences with the other potential. Not surprisingly, sequences with bad z-score values in one potential do not score well in the other one. We observe a strong correlation e.g. between the two potential functions, see Figure 2 and Table 11. However, sequences that are native-like in one potential usually have insufficient z-scores in the other one.

The results of our adaptive walks indicate that the sets $S(\psi)$ are large (indeed, we never encountered the same inverse folded sequence twice) and spread out in sequence space. The data collected in Table 13 show that the average Hamming distances $\langle d \rangle_{\mathrm{adw}}$ of inverse folded sequences are comparable to the sequence length for both potential functions. We find, therefore, that the elements of $S(\psi)$ are approximately randomly distributed over sequence space.

5 The Protein Structures

Due to the nature of the potentials employed for our calculations, only globular proteins were selected for our experiments. Prior to their use, their suitability was established with the `Prosa II` Potential, to ensure, that all their parameters (C_α and C_β z-score, energy values) lie within the desired range. The proteins were also chosen according to their size, the largest protein used contains 129 amino acids (`1lyz`) to ensure that the computer time necessary for each experiment remained within a tolerable limit. The structures used were all well refined and scored well within the `PROSA II` energy ranking. These are the proteins in detail.[2)]

5.1 P22 C2 Repressor

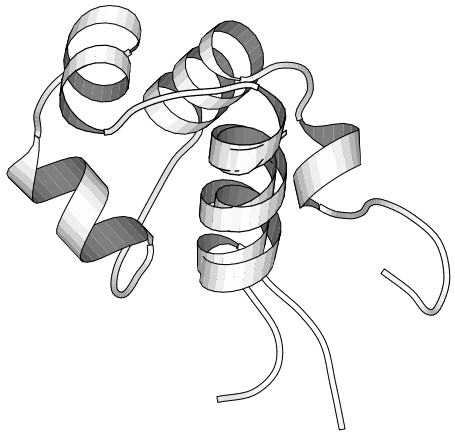

FIGURE 4: DNA-binding domain of P22, PDB Structure `1ADR`

The PDB Structure `1ADR` belongs to the amino-terminal DNA-binding domain, residues 1 - 76 of the P22 c2 Repressor, derived from Salmonella bacteriophage p22 and expressed in (Escherichia coli), Its structure was elucidated by means of NMR studies. This protein allows the phage to reside inactively in the chromosome of its host bacterium. This lysogenic state is maintained by binding of the regulatory protein c2 to the or and ol operators, preventing the transcription of proteins

[2)] All Information about structural motifs were taken from PDBsum:
`http://www.biochem.ucl.ac.uk/bsm/pdbsum/desc.html`

necessary for lytic development. P22 is equivalent and similar to lambda repressor protein c_i.

TABLE 2: Structure Motifs of 1ADR

Motifs	Number	Start	End	n	Sequence
α-helices	1	6	17	12	MGERIRARRKKL
	2	21	28	8	QAALGKMV
	3	32	39	8	NVAISQWE
	4	47	56	10	GENLLALSKA
	5	61	66	6	PDYLLK
β-turns	1	41	44	4	SETE
	2	72	75	4	TNVA
	3	73	76	4	NVAY

The wildtype sequence of the amino-terminal DNA-binding domain is: MNTQLMGERIRARRKKLKIRQAALGKMVGVSNVAISQWERSETEPNGENLLALSKALQCSPDYLLK GDLSQTNVAY. Its principal structural motifs are 5 α-helices and 3 β-turns see Table 2.

5.2 Ubiquitin

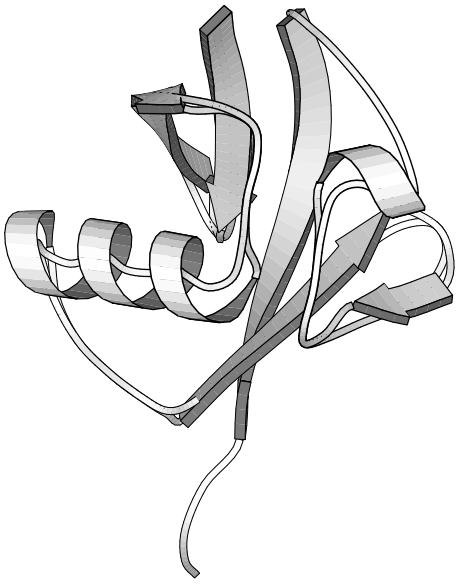

FIGURE 5: Ubiquitin from Human Erythrocytes, PDB Structure 1UBQ

Ubiquitin is a small protein present in all eucaryotic cells, hence the name. 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.

The PDB structure of Ubiquitin from human erythrocytes resolved at Å 1.8 was used for our calculations Its wildtype sequence is: `MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKEST LHLVLRLRGG`. It contains 3 α-helices, 6 β-turns and 5 β-strands see Table 3 This protein structure was already used for extensive prior studies [3], therefore its properties and behaviour in this experiments were already well established.

TABLE 3: Structure Motifs of `1UBQ`

Motifs	Number	Start	End	n	Sequence
α-helices	1	23	34	12	IENVKAKIQDKE
	2	38	40	3	PDQ
	3	57	59	3	SDY
β-turns	1	7	10	4	TLTG
	2	18	21	4	EPSD
	3	44	47	4	IFAG
	4	45	48	4	FAGK
	5	51	54	4	EDGR
	6	62	65	4	QKES
β-strands	1	2	7	6	QIFVKT
	2	12	16	5	TITLE
	3	41	45	5	QRLIF
	4	48	49	2	KQ
	5	66	71	6	TLHLVL

5.3 Calbindin D9K

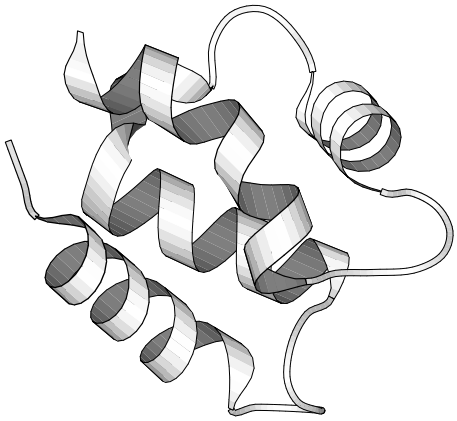

FIGURE 6: Calbindin from Bos Taurus, PDB Structure 4ICB

4ICB is an intestinal vitamin D-dependent calcium-binding protein, it was isolated from Bos Taurus (Bovine) and refined by X-ray diffraction at 1.60Å. It is similar to other ef-hand calcium binding proteins and more specifically to s-100/cabp like proteins. Its wildtype sequence is:
MKSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGEVSFEE
FQVLVKKISQ. The principal structural motifs are 7 α-helices, 3 β- and 1 γ-turn see Table 4.

TABLE 4: Structure Motifs of 4ICB

Motifs	Number	Start	End	n	Sequence
α-helices	1	3	14	12	PEELKGIFEKYA
	2	25	35	11	KEELKLLLQTE
	3	37	40	4	PSLL
	4	46	53	8	LDELFEEL
	5	63	66	4	FEEF
	6	67	69	3	QVL
	7	70	74	5	VKKIS
β-turns	1	17	20	4	EGDP
	2	19	22	4	DPNQ
	3	54	57	4	DKNG
γ-turn	1	53	55	3	LDK

5.4 Thioredoxin

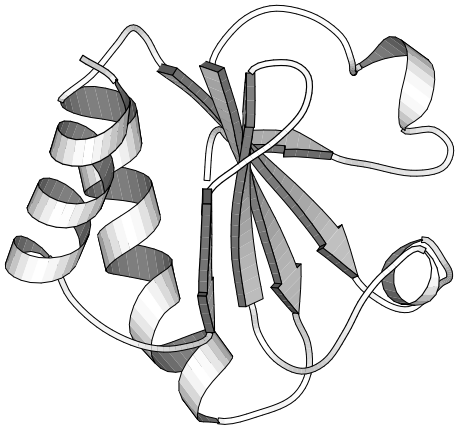

FIGURE 7: Thioredoxin from E. Coli, PDB Structure 2TRXA

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 PDB Structure 2TRXA used for our studies is a thioredoxin from E. Coli (108 amino acids) resolved at 1.68 Å. Its main structural motifs are four α-Helices and one β-sheet consisting of five β-strands see Table 5 and 10 β-turns (not listed in Table 5. The active site is located at the amino-terminus of the second alpha-helix. It contains a disulfide bridge between Cys32 and Cys35. Its wildtype sequence is:
SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGT
APKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLA.
This protein structure was also used for extensive prior studies [3], therefore its properties and behaviour in this experiments were already well established.

The following two proteins were chosen for their globular structure as well as for their sequence length to act as partners for the closest approach studies with Thioredoxin, an other important feature is their structural dissimilarity to thioredoxin.

TABLE 5: Structure Motifs of 2TRXA

Motifs	Number	Start	End	n	Sequence
α-helices	1	12	15	4	FDTD
	2	33	48	16	GPCKMIAPILDEIADE
	3	66	69	4	TAPK
	4	96	107	12	KGQLKEFLDANL
β-strands	1	4	6	3	IIH
	2	22	28	7	AILVDFW
	3	53	59	7	LTVAKLN
	4	77	82	6	TLLLFK
	5	85	91	7	EVAATKV

5.5 Cystatin

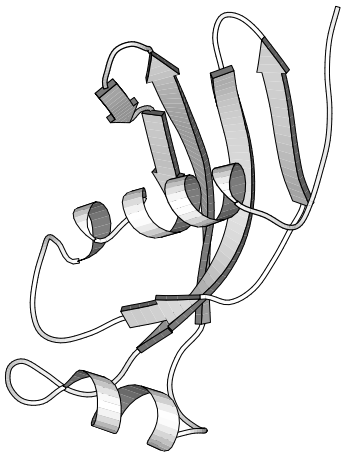

FIGURE 8: Cystatin from hen egg white, PDB Structure 1CEW

Cystatin (1CEW) isolated from hen egg white is a phosphoprotein and a cysteine proteinase inhibitor belonging to the same super-family as the stefin family and the kininogen family [6]. This protein binds tightly to and inhibits a variety of thiol proteases including ficin, papain, and cathepsins b, c, h, and l. Although isolated from egg white, it is also present in serum. The main structural motifs are 1 β-sheet, 3 α-helices and 5 β-strands (see Table 6). Furthermore it contains 2 disulphide bridges. Its wildtype sequence is:
GAPVPVDENDEGLQRALQFAMAEYNRASNDKYSSRVVRVISAKRQLVSGIKYILQVEIGRTTCPKS
SGDLQSCEFHDEPEMAKYTTCTFVVYSIPWLNQIKLLESKCQ.

TABLE 6: Structure Motifs of 1CEW

Motifs	Number	Start	End	n	Sequence
α-helices	1	19	28	10	EGLQRALQFA
	2	30	33	4	AEYN
	3	78	85	8	LQSCEFHD
β-strands	1	12	13	2	VP
	2	40	54	15	YSSRVVRVISAKRQL
	3	58	72	15	IKYILQVEIGRTTCP
	4	92	102	11	YTTCTFVVYSI
	5	107	115	9	QIKLLES

5.6 Rat Oncomodulin

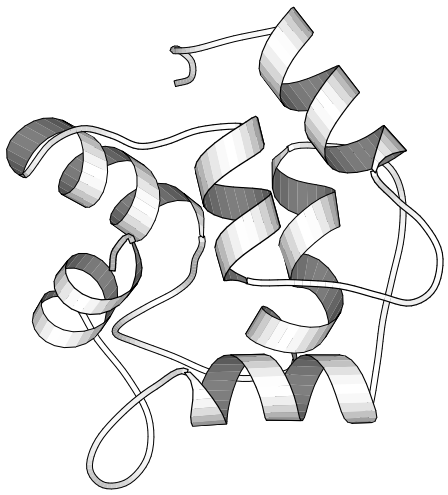

FIGURE 9: Oncomodulin from rat tumours, PDB Structure 1RRO

Oncomodulin belongs to the Parvalbumin sub-family, it is a calcium-binding protein, it has some calmodulin-like activity with respect to enzyme activation and growth regulation and can be found in tumor tissues and not detected in normal tissues. The PDB Structure 1RRO is that of rat oncomodulin, it was isolated from rat tumours (Morris Hepatoma) and resolved at 1.30Å. Its principal structural motifs are 2 β-strands, 9 α-Helices, 5 β-turns (see Table 7) and array of 3 hairpins and two EF-hands (not listed in Table 7). The wildtype sequence is:
SITDILSAEDIAAALQECQDPDTFEPQKFFQTSGLSKMSASQVKDIFRFIDNDQSGYLDGDELKYF
LQKFQSDARELTESETKSLMDAADNDGDGKIGADEFQEMVHS.

TABLE 7: Structure Motifs of 1RRO

Motifs	Number	Start	End	n	Sequence
α-helices	1	2	4	3	ITD
	2	8	17	10	AEDIAAALQE
	3	26	33	8	PQKFFQTS
	4	35	37	3	LSK
	5	40	50	11	ASQVKDIFRFI
	6	61	64	4	DELK
	7	66	69	4	FLQK
	8	79	89	11	ESETKSLMDAA
	9	99	106	8	ADEFQEMV
β-strands	1	57	58	2	YL
	2	97	98	2	IG
β-turns	1	20	23	4	DPDT
	2	51	54	4	DNDQ
	3	69	72	4	KFQS
	4	71	74	4	QSDA
	5	90	93	4	DNDG

5.7 Lysozyme

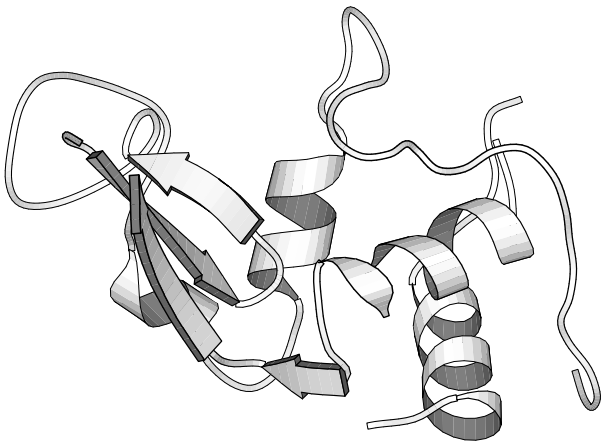

FIGURE 10: Lysozyme from chicken egg white, PDB Structure 1LYZ

Lysozyme is an enzyme capable of dissolving certain bacteria by lysis meaning by cleaving the polysaccharide component of their cell wall. It is a relatively small enzyme. The lysozyme from chicken egg white resolved at 2.0 Åwhich was used for our calculations, is a single polypetide chain of 129 residues. This highly stable protein is cross-linked by four disulfide bridges: between Cys 6 and 127, 30

TABLE 8: Structure Motifs of 1Lyz

Motifs	Number	Start	End	n	Sequence
α-helices	1	5	14	10	RCELAAAMKR
	2	26	36	11	GNWVCAAKFES
	3	80	84	5	CSALL
	4	89	98	10	TASVNCAKKI
	5	104	107	4	GMNA
	6	109	114	6	VAWRNR
	7	120	123	4	VQAW
β-strands	1	43	45	3	TNR
	2	51	53	3	TDY
	3	58	59	2	IN

and 115, 64 and 80, 76 and 94. The active site contains Asp 52 and Glu 35. Its wildtype sequence is:
KKLGRCELAAAMKRHGLQNERGLSMGNWVCAAAFESNFNTQATNRNTDGSTDYTFLQINSRWWC
NDGRAPGSRNLCGIPCSALLSSDITASVNCAVKIYSDGNGCNIMVAWRNRCKGTDEQRWIRGCRL.
Its principal structural motifs are seven α-Helices, 1 β-sheet consisting of three β-strands (see Table 8) and 13 Turns (not listed in Table 8).

5.8 The Janus Proteins

Cols*E1 Repressor of Primer and Protein G

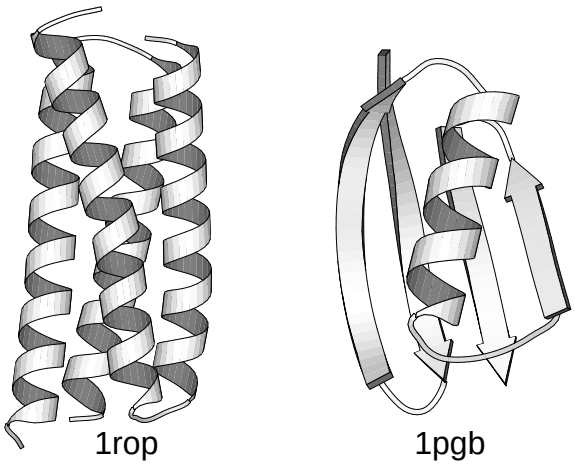

FIGURE 11: PDB Structure 1ROP and 1PGB

TABLE 9: Structure Motifs of 1ROP

Motifs	Number	Start	End	n	Sequence
α-helices	1	3	28	26	KQEKTALNMARF IRSQTLTLLEKLNE
	2	32	55	24	DEQADICESLHDHADELYRSCLAR
γ-turn	1	30	32	3	DAD

The PDB Structure 1ROP is that of the regulatory protein of E.Coli refined at 1,70Å, it regulates plasmid DNA replication by modulating the initiation of transcription of the primer RNA precursor. processing of the precursor of the primer, RNA II, is inhibited by hydrogen bonding of RNA II to its complementary sequence in RNAI. ROP increases the affinity of RNAI for RNA II and thus decreases the rate of replication initiation events. In its native form it is an anti-parallel homo dimer. A monomer of 1rop contains 2 α-helices and 1 γ-turn (see Table 9). Its wildtype sequence is :

MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARFGDDGENL

TABLE 10: Structure Motifs of 1PGB

Motifs	Number	Start	End	n	Sequence
α-helix	1	23	36	14	AATAEKVFKQYAND
β-strands	1	2	8	7	TYKLILN
	2	13	19	7	KGETTTE
	3	42	46	5	EWTYD
	4	51	55	5	TFTVT
β-turnss	1	9	12	4	GKTL
	2	46	49	4	DDAT
	3	47	50	4	DATK

The PDB-Structure 1PGB belongs to the b1 igG-binding domain of Protein G isolated from Streptococcus, Lancefield group G. It was resolved at 1.92 Å. Protein G is a small globular protein produced by several Streptococcal species. Protein G's bind the Fc regions of IgG very tightly, in this functional characteristic, they resemble the staphylococcal protein A Its main structural motifs are 1β-sheet consisting of 4 β-strands, 1 α-helix, 3 beta turns (see Table 10) and 2 β-hairpins (not listed in Table 10). Its wildtype sequence is:

MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

6 Adaptive walks

As mentioned in section 4.3 adaptive walks yield insight into the structure of protein space. The length ℓ of adaptive walks gives information about the ruggedness of the energy landscape [38]. Longer walks imply smoother surfaces with few local optima. In this section we discuss in detail the results of our adaptive walk experiments.

In order to compare the predictions from both potentials we have taken adaptive walks computed with one potential and re-evaluated the sequences with the other potential. As would be anticipated, sequences with bad z-score values in one potential do not score well in the other one. Thus, we observe a strong correlation between the two potential functions. Still, as a rule we have found that sequences that are native-like in one potential usually have insufficient z-scores in the other one.

Adaptive walks can be used to optimize the z-score of a sequence well beyond the native-like threshold level z^*. We find that sequences with unnaturally good z-score levels in one potential often have z-scores at least close to the native value of the other potential. It is interesting to note that sequences optimized with the `NN` potential yield more native-like PROSA z-scores than *vice versa*. In the following sections, we discuss the results of the re-evaluation of the adaptive walks in detail.

Note that better z-scores in the `NN` Potential are more positive (see e.g. Figure 22), while in the `PROSA II` Potential better z-scores have more negative values (see e.g. Figure 12).

6.1 Adaptive Walks `1ADR`

The protein `1ADR` ($n = 76$) is the only protein that was only evaluated in one potential `PROSA II`. It was impossible to integrate it into the fingerprint database of the `NN` Potential. Hence, the results for this protein could not be re-evaluated with the other potential. Figure 12 shows the results of the 10 adaptive walks

with the PROSA II potential. The algorithm terminated each run at a predefined threshold score z^*. We choose z^* 10 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -7.02$). In the case of the PROSA II potentials we require that both the C^α and the C^β z-scores improve with each step of the adaptive walk. For 1ADR we found that the average length ℓ of all adaptive walks to reach wildtype z-score was 31.7, corresponding to 41.7 % of the sequence length.

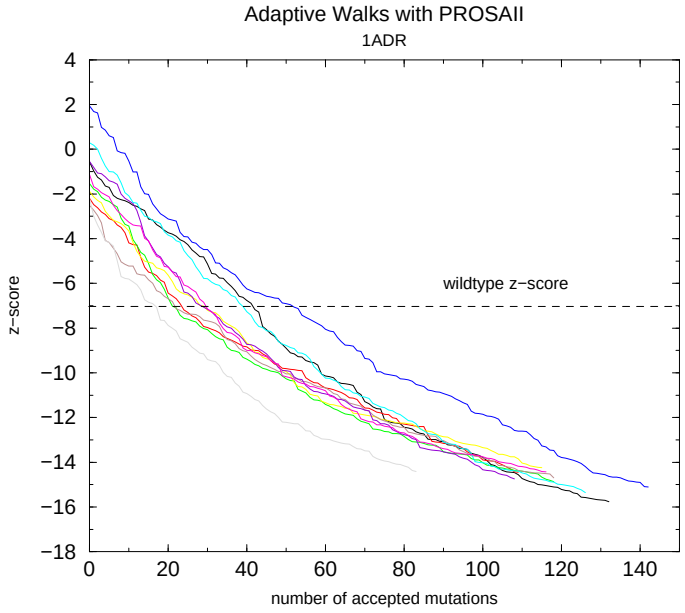

FIGURE 12: 1ADR, 10 Adaptive Walks with PROSA II

6.2 Adaptive Walks 1UBQ

1UBQ ($n = 76$) was evaluated with both the PROSA II and the NN Potential.

Figure 13 shows the results of the 10 adaptive walks with the PROSA II potential. The algorithm terminated each run at a predefined threshold score z^*, z^* being 6 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -9.26$). We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 61.9, corresponding to 81.4 % of the sequence length.

The data from these runs was re-evaluated using the NN Potential (see Figure 14).

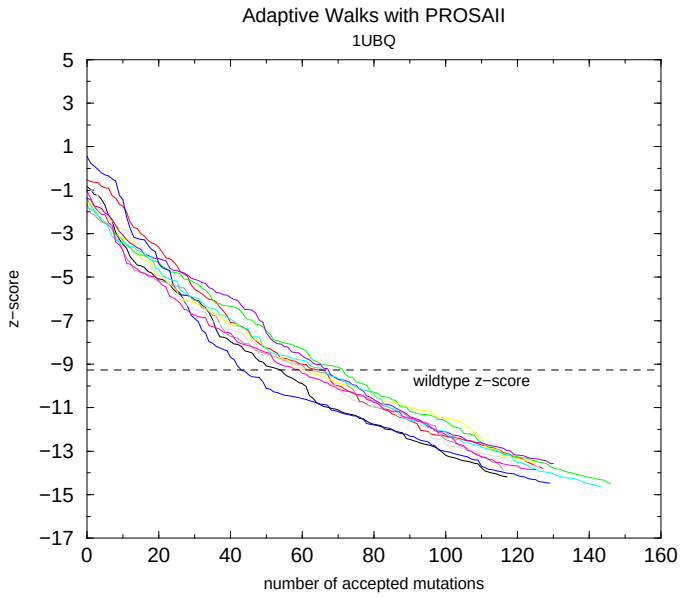

FIGURE 13: 1UBQ, 10 Adaptive Walks with `PROSA II`

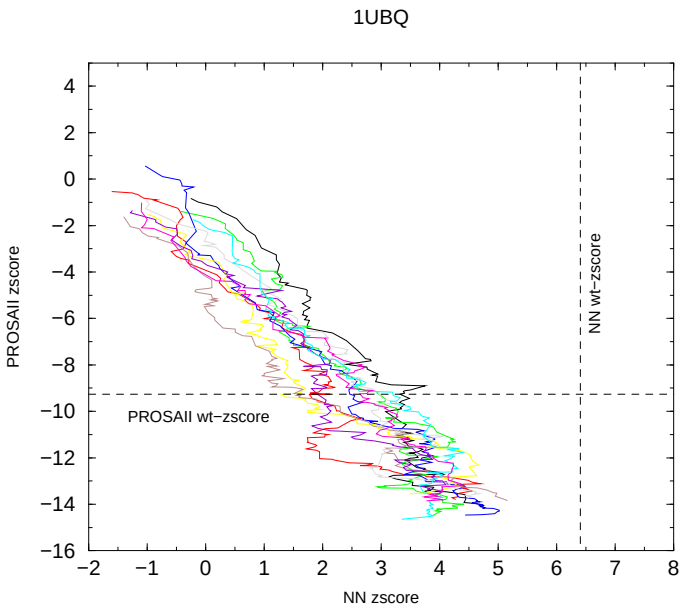

FIGURE 14: 1UBQ, 10 Adaptive Walks with `PROSA II` evaluated with the `NN` Potential

Obviously even the sequences that scored well above wildtype level in the `PROSA II` potential do not reach wildtype level in the `NN` Potential. The best z-score that was achieved in the `NN` Potential was 5.14 which is approximately one units below the `NN` z-score for wildtype `1UBQ`. However, calculating the regression of

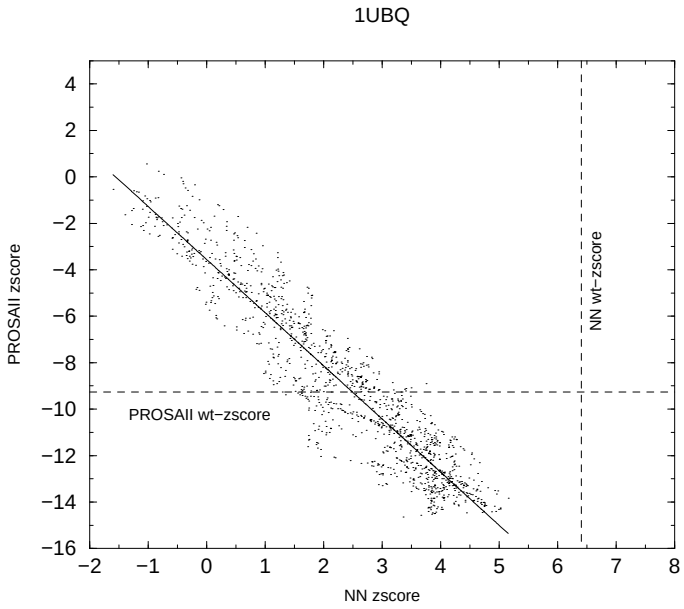

FIGURE 15: Regression of 10 Adaptive Walks in Figure 14

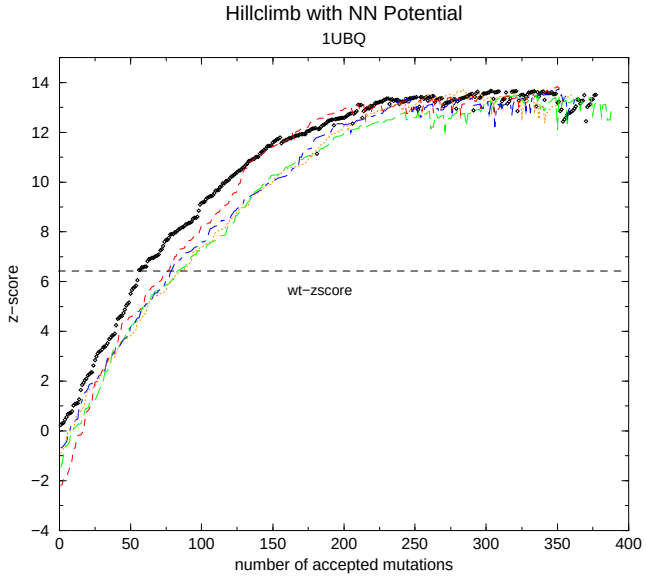

FIGURE 16: 1UBQ, 5 Adaptive Walks with the `NN` Potential

the results from this comparison (see Figure 15), we find that along the adaptive walks, we obtain an approximately linear relationship.

Figure 16 shows the results of the 5 adaptive walks with the `NN` potential. No predefined threshold z-score was implemented to terminate the runs for the `NN`

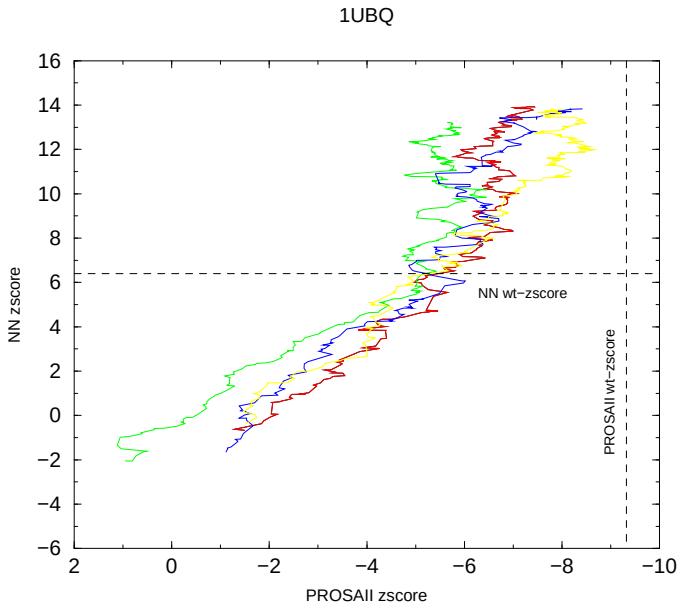

FIGURE 17: 1UBQ, 5 Adaptive Walks with the `NN` Potential evaluated with `PROSA II`

adaptive walks. The modus of these runs was to mutate a sequence, accepting only those changes that don't decrease the fitness in comparison to the previous test sequence. The z-scores evolved well beyond the z-score of the wildtype sequence/structure pair ($z_{\text{wt}} = 6.40$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 75.4, corresponding to 99.2 % of the sequence length.

The data from these runs was re-evaluated using the `PROSA II` Potential (see Figure 17). The sequences that scored well above wildtype level in the `NN` potential did not reach wildtype level in the `PROSA II` potential. The best z-score that was achieved in the `PROSA II` Potential by the `NN` sequences, was 8.7 which is 0.5 units below the `PROSA II` z-score for wildtype `1UBQ`. Calculating the regression of the results from this comparison (see Figure 18), we again obtain an approximately linear relationship.

6.3 Adaptive Walks `4ICB`

`4ICB` ($n = 76$) was evaluated with both the `PROSA II` and the `NN` Potential.

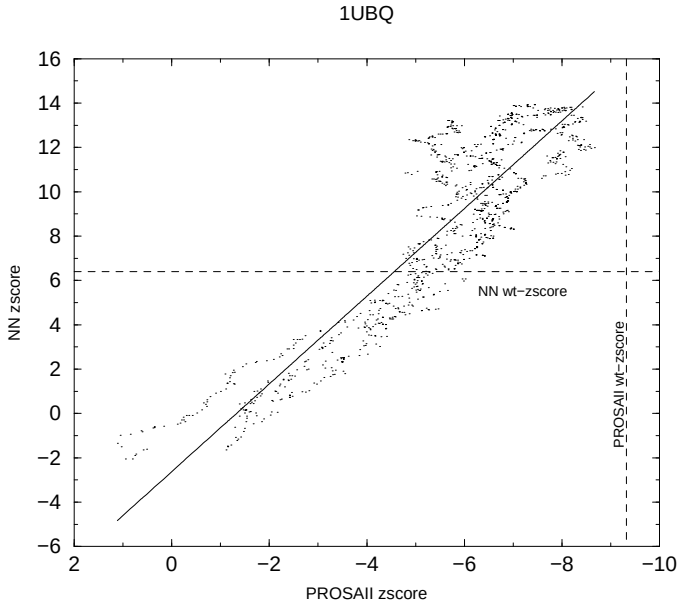

FIGURE 18: Regression of 5 Adaptive Walks in Figure 17

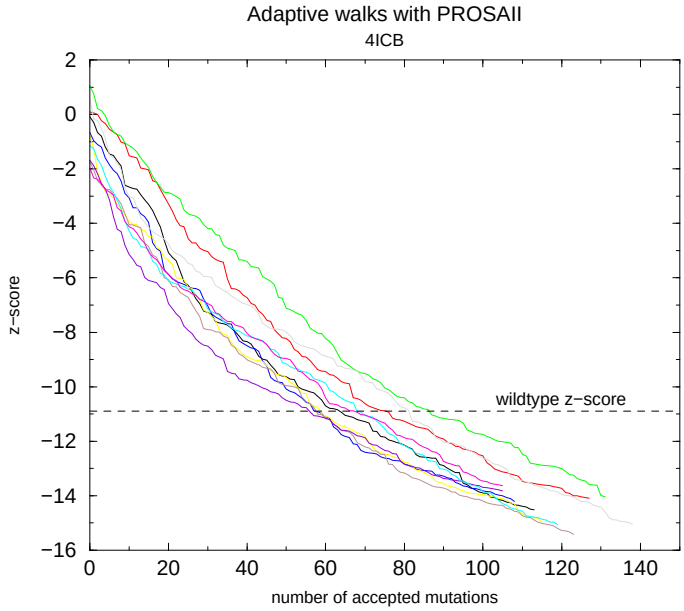

FIGURE 19: 4ICB, 10 Adaptive Walks with `PROSA II`

Figure 19 shows the results of the 10 adaptive walks with the `PROSA II` potential. The algorithm terminated each run at a predefined threshold score z^*, z^* being 5 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -8.08$).

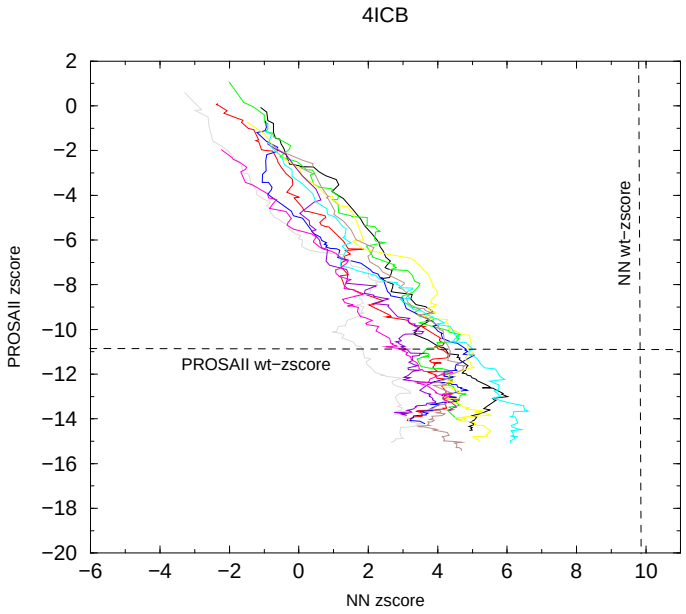

FIGURE 20: 4ICB, 10 Adaptive Walks with `PROSA II` evaluated with the `NN` Potential

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 60.3, corresponding to 79.3 % of the sequence length.

The data from these runs was re-evaluated using the `NN` Potential (see Figure 20). Obviously even the sequences that scored well above wildtype level in the `PROSA II` potential do not reach wildtype level in the `NN` Potential.

The best z-score that was achieved in the `NN` Potential was 6.6 which is approximately two units below the `NN` z-score for wildtype `4ICB`. However, calculating the regression of the results from this comparison (see Figure 21), we find that along the adaptive walks, we obtain an approximately linear relationship.

Figure 22 shows the results of the 5 adaptive walks with the `NN` potential. No predefined threshold z-score was implemented to terminate the runs for the `NN` adaptive walks. The modus of these runs was to mutate a sequence, accepting only those changes that don't decrease the fitness in comparison to the previous test sequence. The z-scores evolved well beyond the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = 8.08$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 75.4, corresponding to 99.2 % of the sequence length.

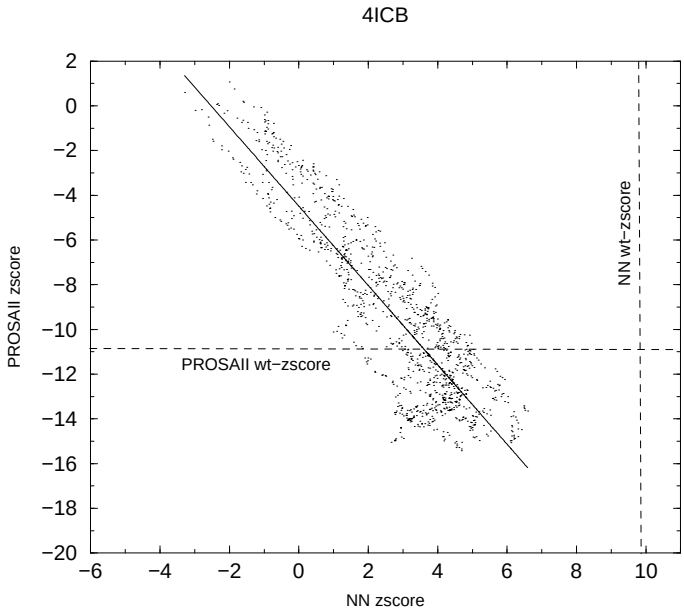

FIGURE 21: Regression of 10 Adaptive Walks in Figure 20

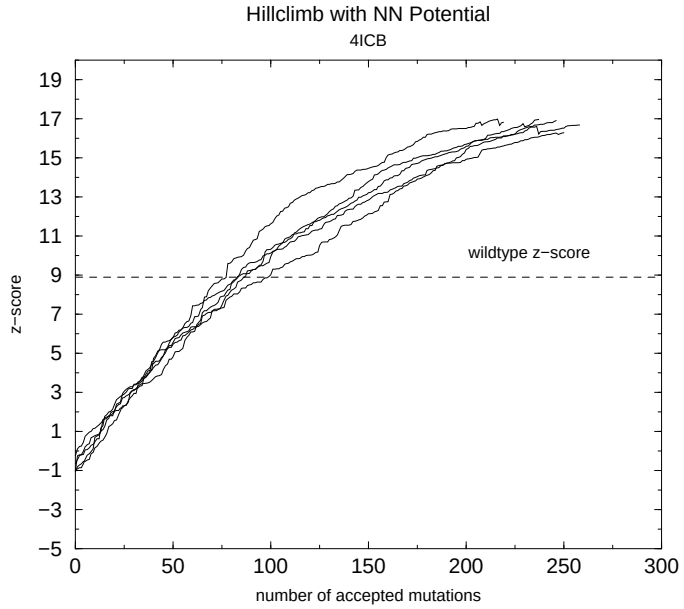

FIGURE 22: 4ICB, 5 Adaptive Walks with the `NN` Potential

The data from these runs was re-evaluated using the `PROSA II` Potential (see Figure 23). The sequences that scored well above wildtype level in the `NN` potential did not reach wildtype level in the `PROSA II` Potential, the best z-score that was achieved in the `PROSA II` Potential by the `NN` sequences, was 8.1 which is two units

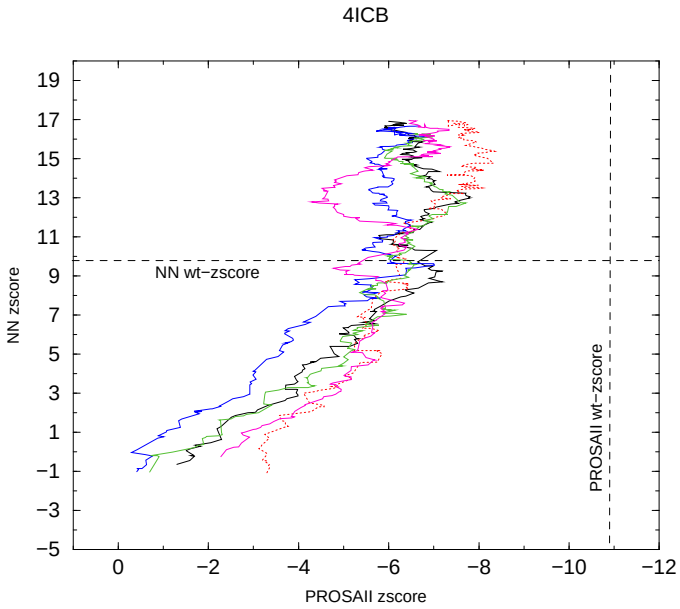

FIGURE 23: 4ICB, 5 Adaptive Walks with the `NN` Potential evaluated with `PROSA II`

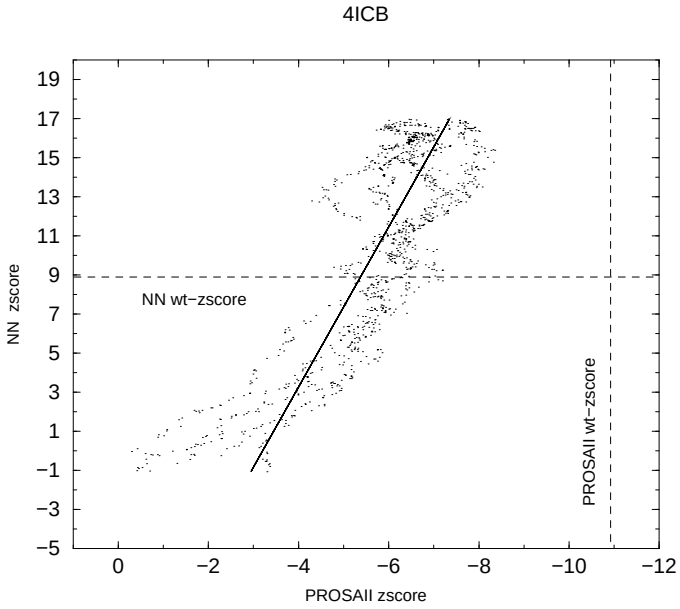

FIGURE 24: Regression of 5 Adaptive Walks in Figure 23

below the `PROSA II` z-score for wildtype `4ICB`. Calculating the regression of the results from this comparison (see Figure 24), we again obtain an approximately linear relationship.

6.4 Adaptive Walks 1CEW

1CEW ($n = 108$) was evaluated with both the PROSA II and the NN Potential. This was the only case were both the sequences from the PROSA II adaptive walks and the NN adaptive walks reached above wildtype level when they were re-evaluated in the other potential!

Figure 25 shows the results of the 10 adaptive walks with the PROSA II potential.

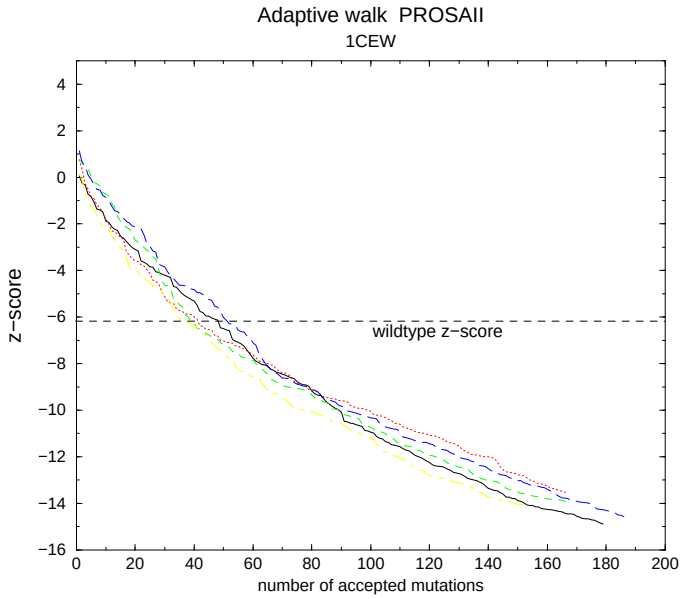

FIGURE 25: 1CEW, 10 Adaptive Walks with PROSA II

The algorithm terminated each run at a predefined threshold score z^*, z^* being 10 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -5.91$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 44.1, corresponding to 40,8 % of the sequence length. The data from these runs was re-evaluated using the NN Potential (see Figure 26). Obviously even the sequences that scored well above wildtype level in the PROSA II barely reached wildtype level in the NN Potential, the best z-score that was achieved in the NN Potential was 6.9 which is slightly better than the NN z-score for wildtype 1CEW. Calculating the regression of the results from this comparison (see Figure 21), we find that along the adaptive walks, we obtain an approximately linear

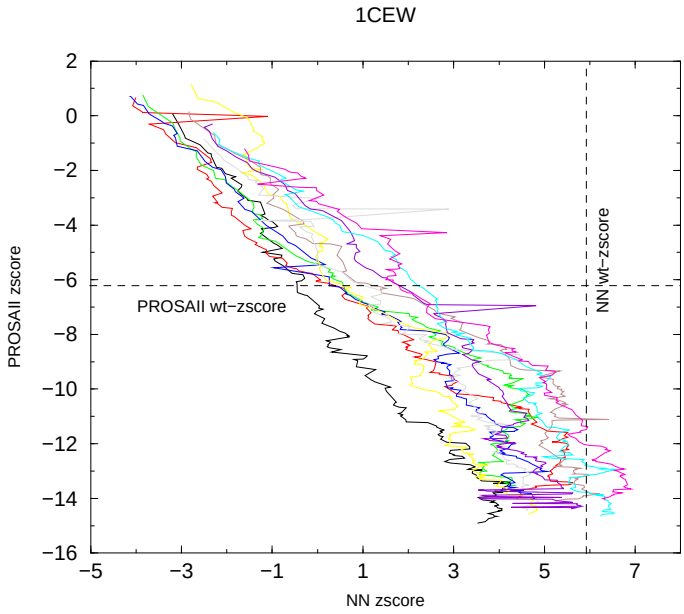

FIGURE 26: 1CEW, 10 Adaptive Walks with `PROSA II` evaluated with the `NN` Potential

relationship.

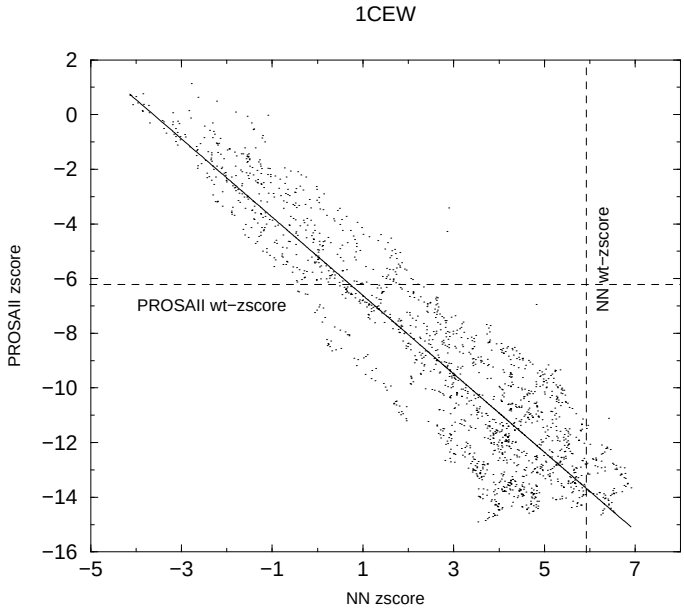

FIGURE 27: Regression of 10 Adaptive Walks in Figure 26

Figure 28 shows the results of the 5 adaptive walks with the `NN` potential. No predefined threshold z-score was implemented to terminate the runs for the `NN`

adaptive walks.

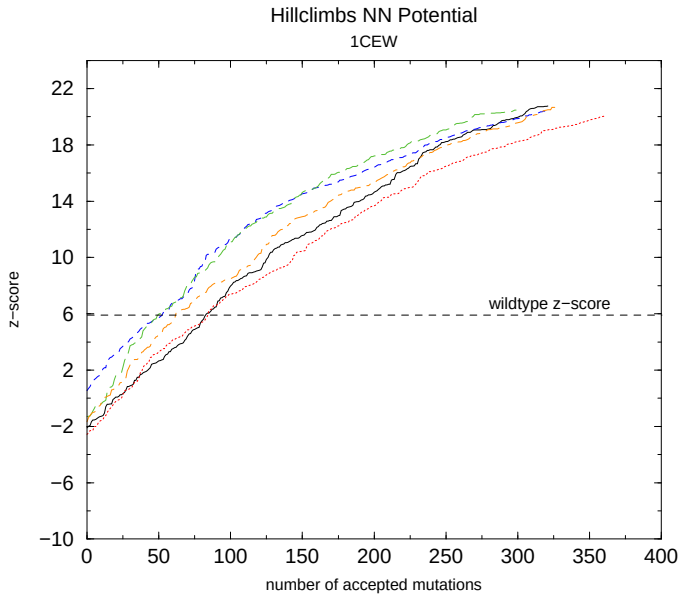

FIGURE 28: 1CEW, 5 Adaptive Walks with the `NN` Potential

The modus of these runs was to mutate a sequence, accepting only those changes that don't decrease the fitness in comparison to the previous test sequence. The z-scores evolved well beyond the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = 6.20$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 76.0, corresponding to 70.3 % of the sequence length.

The data from these runs was re-evaluated using the `PROSA II` Potential (see Figure 29). The sequences that scored well above wildtype level in the `NN` potential reached wildtype level in the `PROSA II` Potential, the best z-score that was achieved in the `PROSA II` Potential by the `NN` sequences, was 6.9 which is one unit above the `PROSA II` z-score for wildtype `1CEW`.

Calculating the regression of the results from this comparison (see Figure 30), we again obtain an approximately linear relationship.

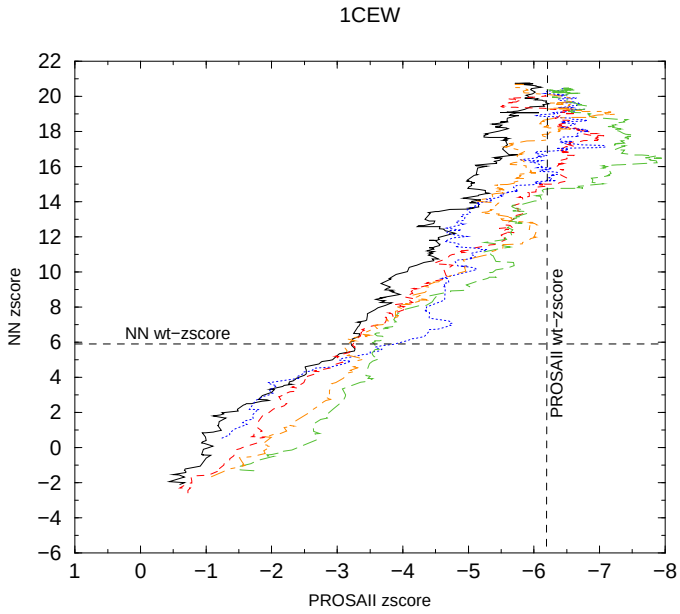

FIGURE 29: 1CEW, 5 Adaptive Walks with the `NN` Potential evaluated with `PROSA II`

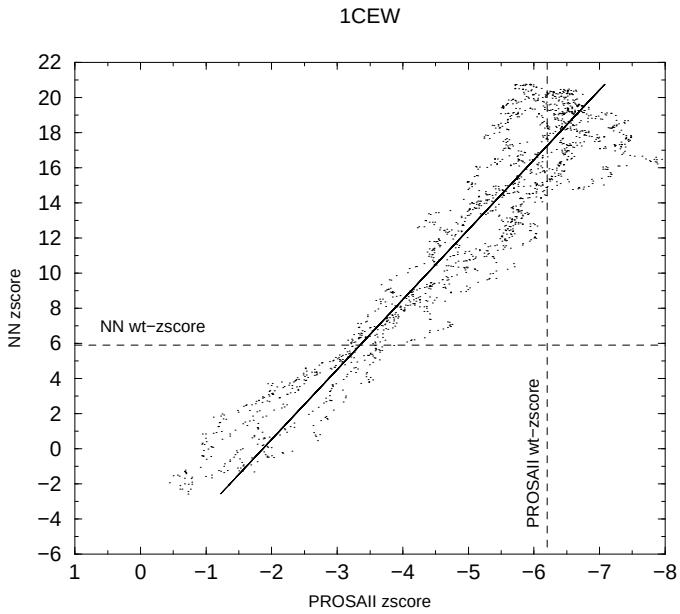

FIGURE 30: Regression of 5 Adaptive Walks in Figure 29

6.5 Adaptive Walks `1RRO`

`1RRO` ($n = 108$) was evaluated with both the `PROSA II` and the `NN` Potential.

Figure 31 shows the results of the 10 adaptive walks with the `PROSA II` potential.

The algorithm terminated each run at a predefined threshold score z^*, z^* being 5 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -10.88$).

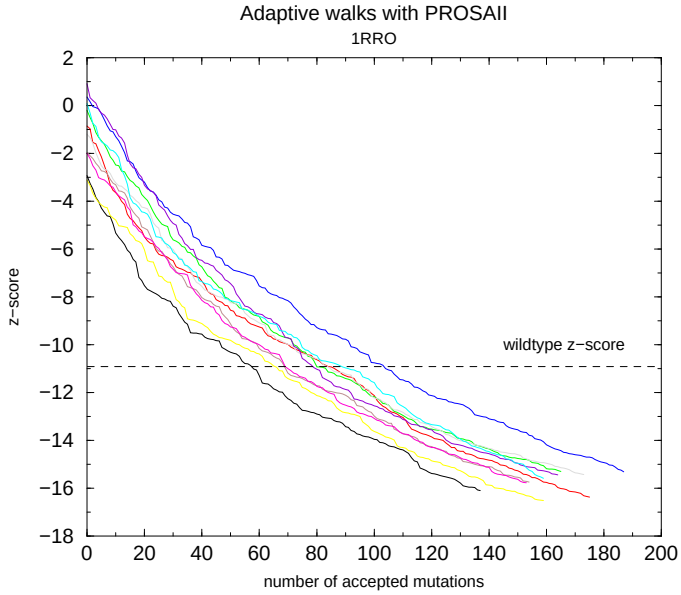

FIGURE 31: 1RRO, 10 Adaptive Walks with PROSA II

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 79.1, corresponding to 73.2 % of the sequence length.

The data from these runs was re-evaluated using the NN Potential (see Figure 32). Obviously even the sequences that scored well above wildtype level in the PROSA II potential do not reach wildtype level in the NN Potential.

The best z-score that was achieved in the NN Potential was 7.2 which is approximately two units below the NN z-score for wildtype 1RRO. However, calculating the regression of the results from this comparison (see Figure 33), we find that along the adaptive walks, we obtain an approximately linear relationship.

Figure 34 shows the results of the 5 adaptive walks with the NN potential. No predefined threshold z-score was implemented to terminate the runs for the NN adaptive walks. The modus of these runs was to mutate a sequence, accepting only those changes that don't decrease the fitness in comparison to the previous test sequence. The z-scores evolved well beyond the z-score of the wildtype se-

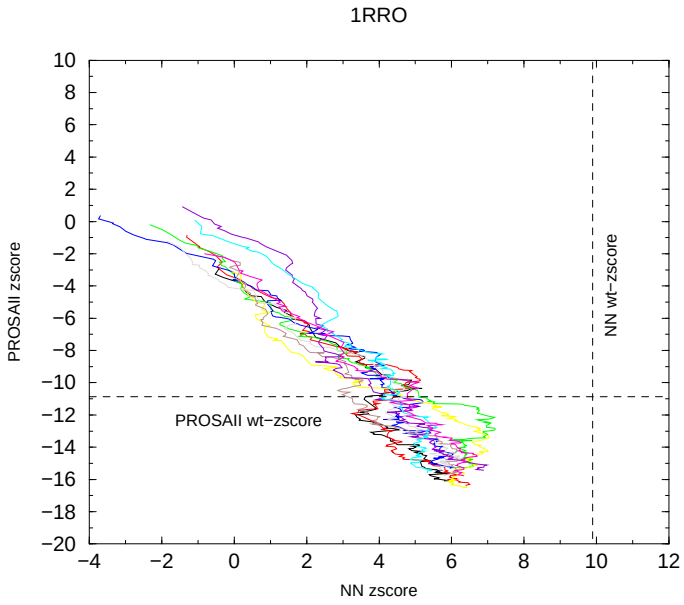

FIGURE 32: 1RRO, 10 Adaptive Walks with PROSA II evaluated with the NN Potential

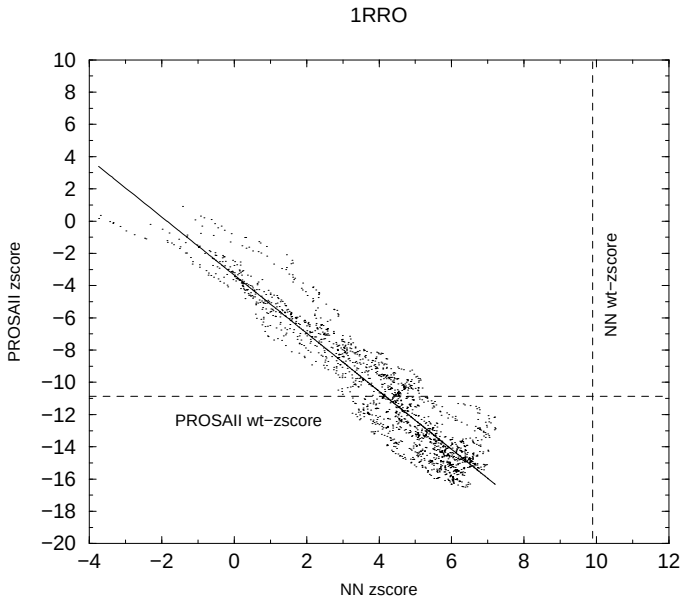

FIGURE 33: Regression of 10 Adaptive Walks in Figure 32

quence/structure pair ($z_{\mathrm{wt}} = 9.80$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 125.6, corresponding to 116.3 % of the sequence length. The data from these runs was re-evaluated using the PROSA II Potential (see Figure 35).

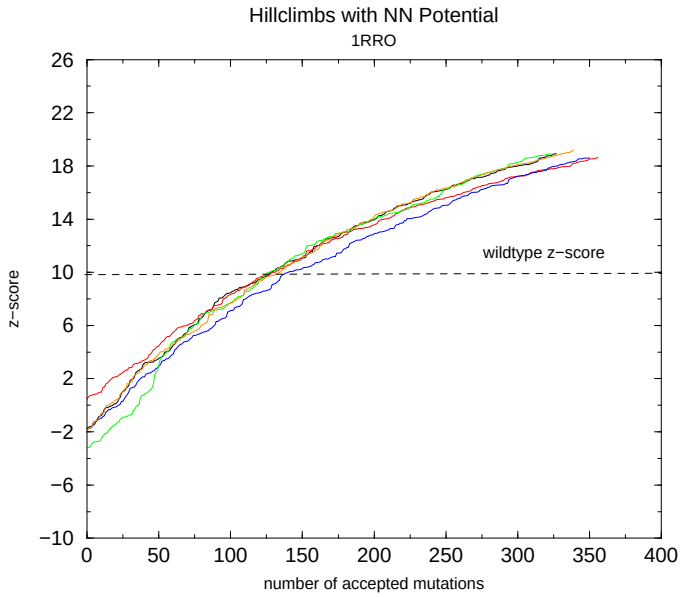

FIGURE 34: 1RRO, 5 Adaptive Walks with the NN Potential

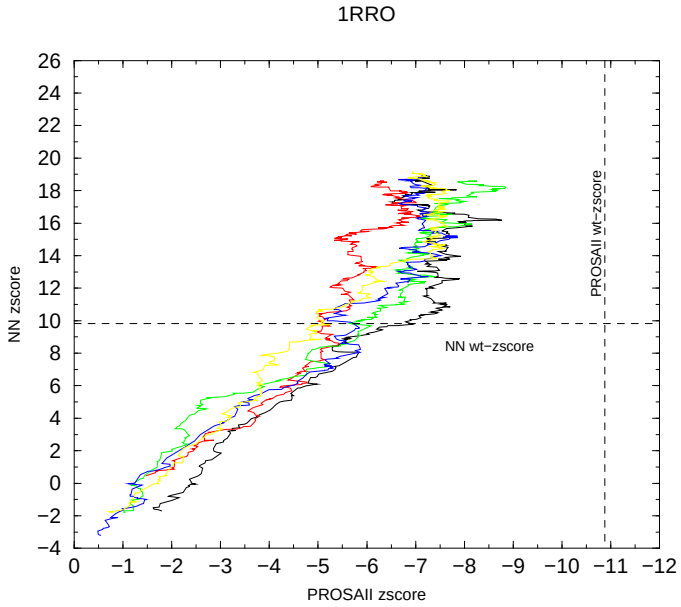

FIGURE 35: 1RRO, 5 Adaptive Walks with the NN Potential evaluated with PROSA II

The sequences that scored well above wildtype level in the NN potential did not reach wildtype level in the PROSA II Potential, the best z-score that was achieved in the PROSA II Potential by the NN sequences, was 8.8 which is two units below the PROSA II z-score for wildtype 1RRO.

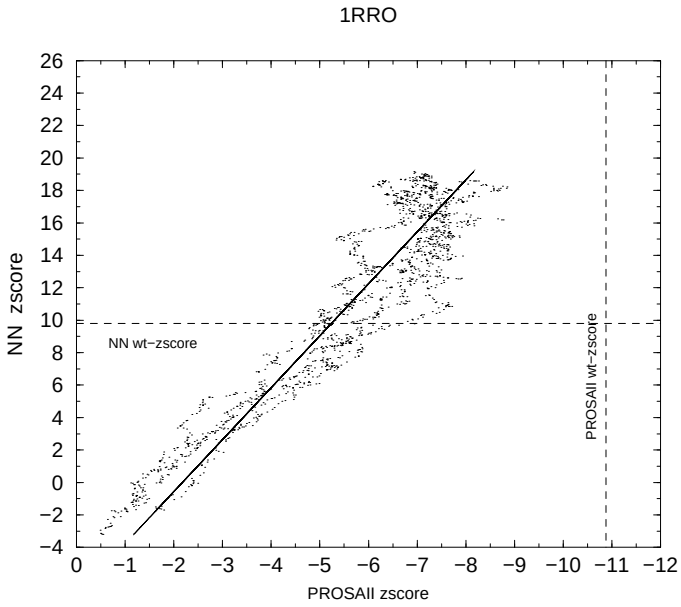

FIGURE 36: Regression of 5 Adaptive Walks in Figure 35

Calculating the regression of the results from this comparison (see Figure 36), we again obtain an approximately linear relationship.

6.6 Adaptive Walks 2TRXA

2TRXA ($n = 108$) was evaluated with both the PROSA II and the NN Potential.

Figure 37 shows the results of the 10 adaptive walks with the PROSA II potential. The algorithm terminated each run at a predefined threshold score z^*, z^* being 6 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -9.22$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 71.7, corresponding to 66.4 % of the sequence length.

The data from these runs was re-evaluated using the NN Potential (see Figure 38). Obviously even the sequences that scored well above wildtype level in the PROSA II potential do not reach wildtype level in the NN Potential.

The best z-score that was achieved in the NN Potential was 6.6 which is approxi-

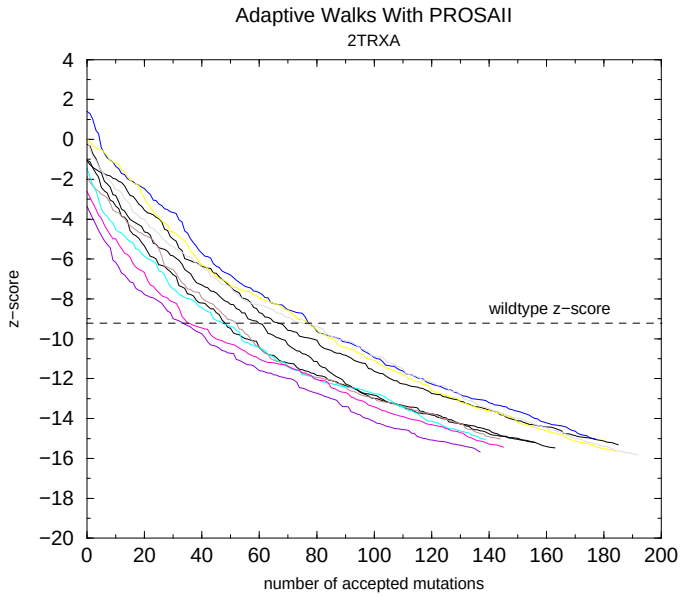

FIGURE 37: 2TRXA, 10 Adaptive Walks with PROSA II

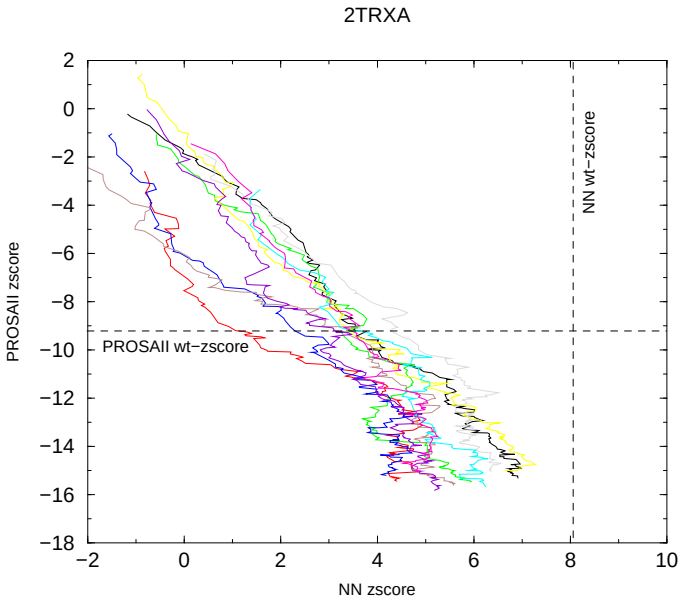

FIGURE 38: 2TRXA, 10 Adaptive Walks with PROSA II evaluated with the NN Potential

mately two units below the NN z-score for wildtype 2TRXA. However, calculating the regression of the results from this comparison (see Figure 39), we find that along the adaptive walks, we obtain an approximately linear relationship.

Figure 40 shows the results of the 5 adaptive walks with the NN potential. No

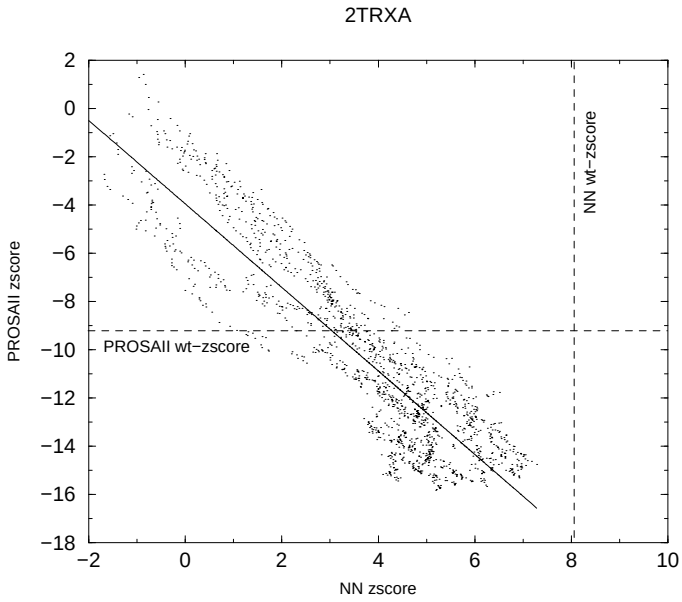

FIGURE 39: Regression of 10 Adaptive Walks in Figure 38

predefined threshold z-score was implemented to terminate the runs for the `NN` adaptive walks.
The modus of these runs was to mutate a sequence, accepting only those changes that don't decrease the fitness in comparison to the previous test sequence. The z-scores evolved well beyond the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = 8.06$).

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 112.0, corresponding to 103.7 % of the sequence length.

The data from these runs was re-evaluated using the `PROSA II` Potential (see Figure 41). The sequences that scored well above wildtype level in the `NN` potential came close to wildtype level in the `PROSA II` Potential.

The best z-score that was achieved in the `PROSA II` Potential by the `NN` sequences, was 7.3 which is 0.8 units below the `PROSA II` z-score for wildtype `2TRXA`. Calculating the regression of the results from this comparison (see Figure 42), we again obtain an approximately linear relationship.

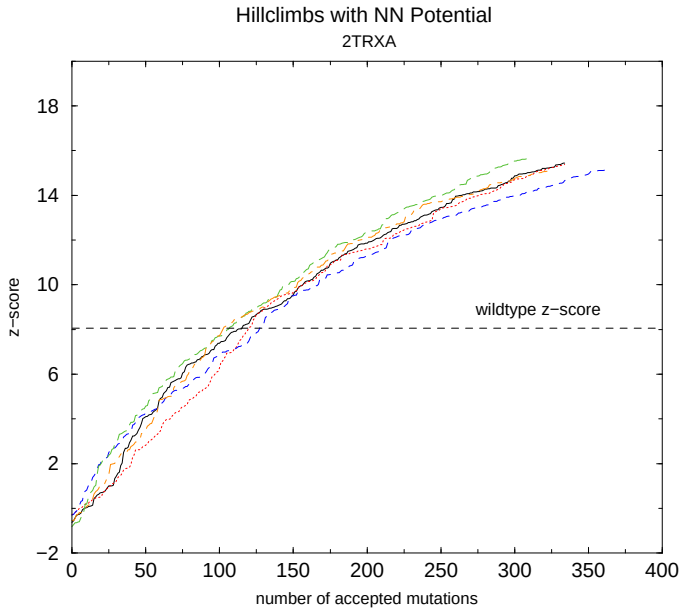

FIGURE 40: 2TRXA, 5 Adaptive Walks with the NN Potential

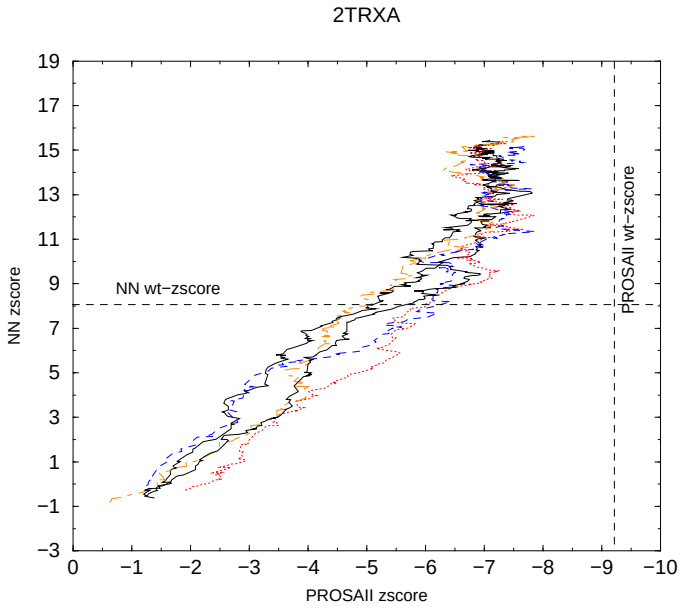

FIGURE 41: 2TRXA, 5 Adaptive Walks with the NN Potential evaluated with PROSA II

6.7 Adaptive Walks 1LYZ

1LYZ ($n = 129$) is the largest protein used in this study, it was evaluated with both the PROSA II and the NN Potential.

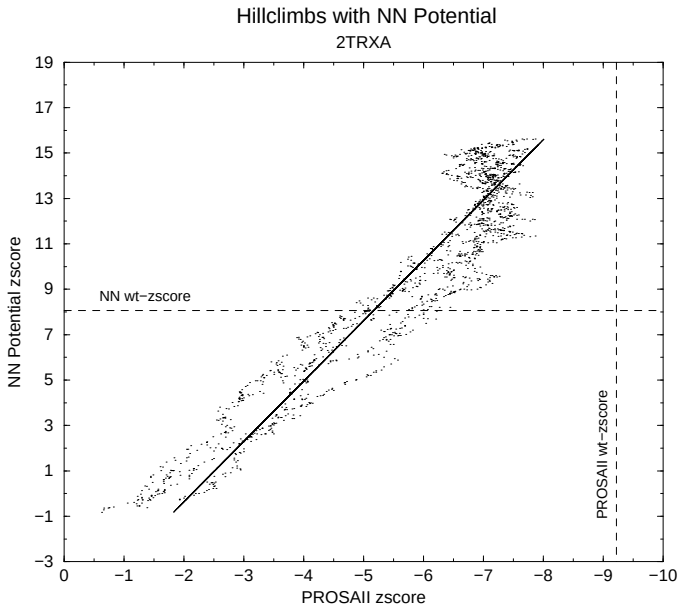

FIGURE 42: Regression of 5 Adaptive Walks in Figure 41

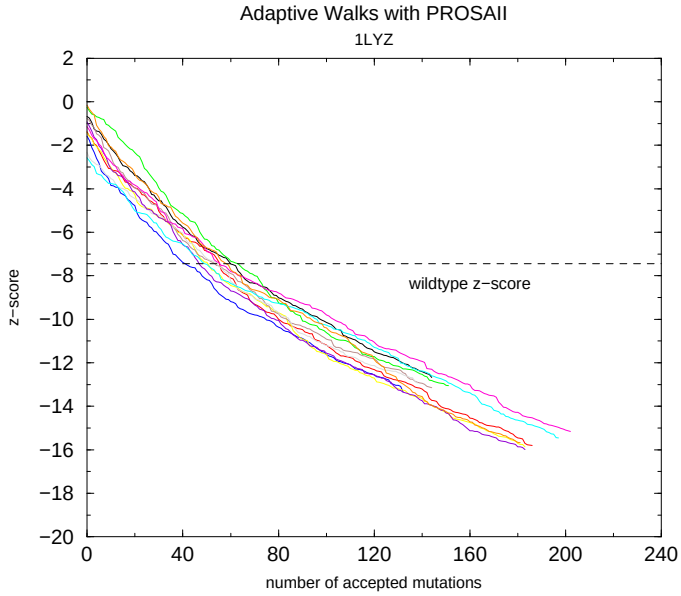

FIGURE 43: 1LYZ, 10 Adaptive Walks with `PROSA II`

Figure 43 shows the results of the 10 adaptive walks with the `PROSA II` potential. The algorithm terminated each run at a predefined threshold score z^*, z^* being 8 z-score units better than the z-score of the wildtype sequence/structure pair ($z_{\mathrm{wt}} = -7.45$).

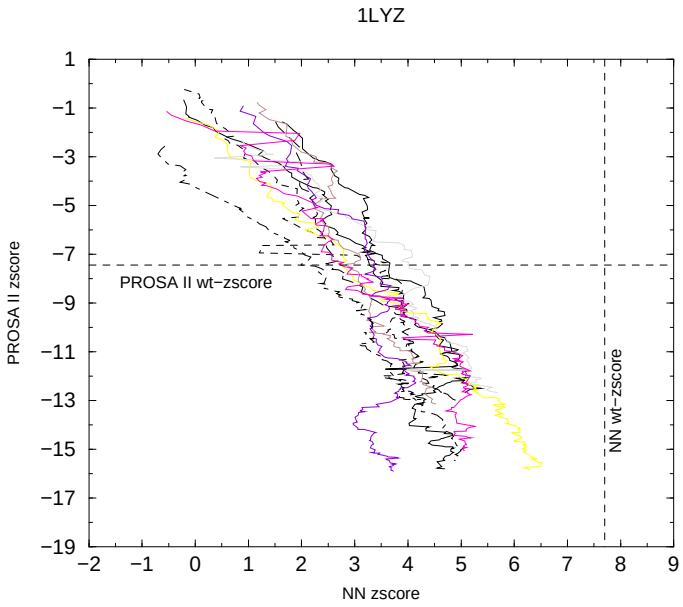

FIGURE 44: 1LYZ, 10 Adaptive Walks with `PROSA II` evaluated with the `NN` Potential

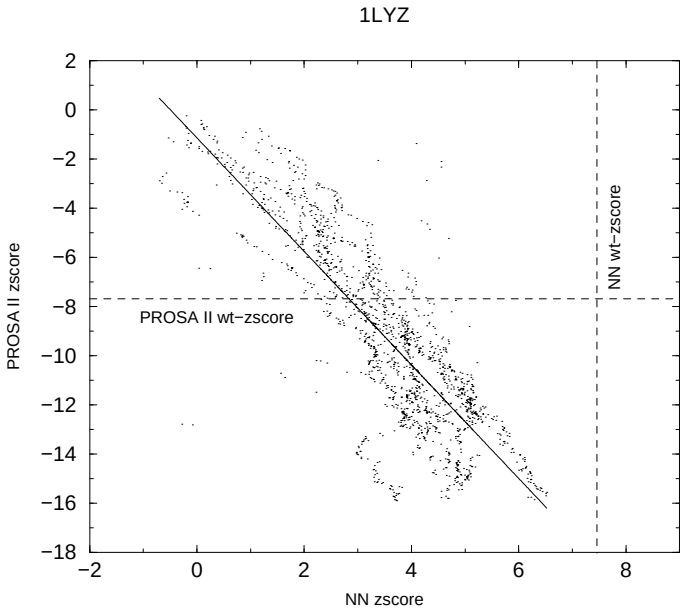

FIGURE 45: Regression of 10 Adaptive Walks in Figure 44

The data from these runs was re-evaluated using the `NN` Potential (see Figure 44). Obviously even the sequences that scored well above wildtype level in the `PROSA II` potential do not reach wildtype level in the `NN` Potential.

The best z-score that was achieved in the `NN` Potential was 6.5 which is appro-

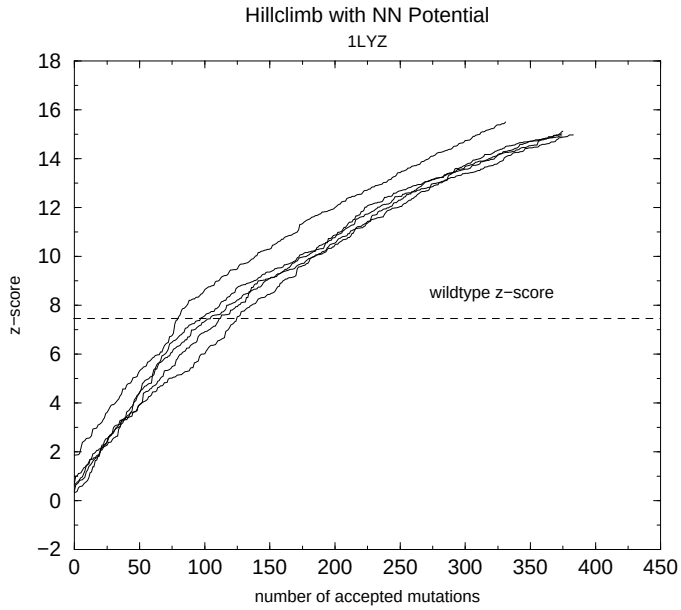

FIGURE 46: 1LYZ, 5 Adaptive Walks with the NN Potential

ximately one units below the NN z-score for wildtype 1LYZ. However, calculating the regression of the results from this comparison (see Figure 45), we find that along the adaptive walks, we obtain an approximately linear relationship.

We found that the average length ℓ of all adaptive walks to reach wildtype z-score was 115.2, corresponding to 89.3 % of the sequence length.

The data from these runs was re-evaluated using the PROSA II Potential (see Figure 47). Again, the sequences generated with the NN Potential scored better in the PROSA II potential, i.e. the sequences that scored well above wildtype level in the NN potential came close to wildtype level in the PROSA II Potential, the best z-score achieved in the PROSA II Potential by theNN sequences, was 7.7 which is only 0.2 units below the PROSA II z-score for wildtype 1LYZ. Calculating the regression of the results from this comparison (see Figure 48), we again obtain an approximately linear relationship.

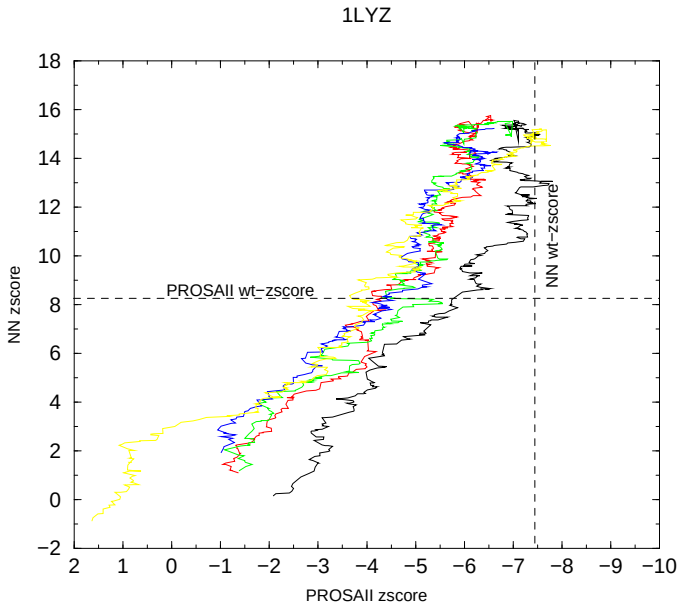

FIGURE 47: 1LYZ, 5 Adaptive Walks with the `NN` Potential evaluated with `PROSA II`

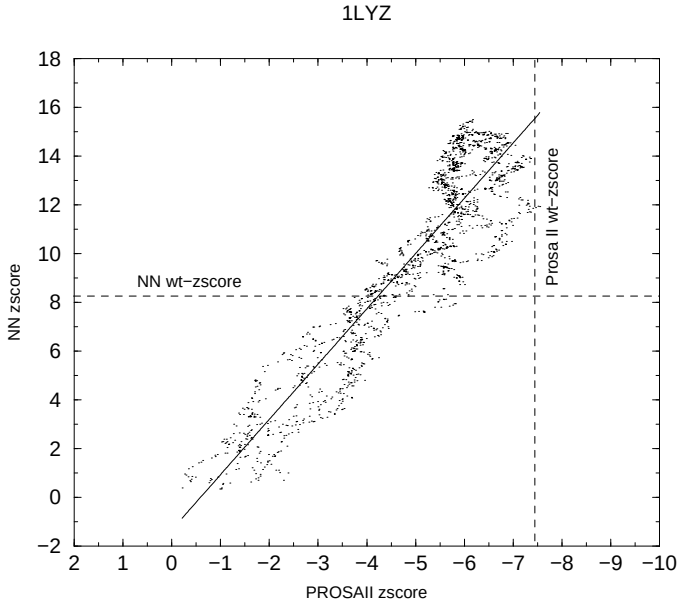

FIGURE 48: Regression of 5 Adaptive Walks in Figure 47

6.8 Summary

We have found that along an adaptive walk, we obtain an approximately linear relationship between the scores from the two potentials (see Table 11).

$$z' = az + b \tag{6}$$

where z is the z-score w.r.t. the potential that is used to optimize the sequences and z' is the z-score w.r.t. the other potential. For each protein, a and b can be determined rather accurately from 5 to 10 independent adaptive walks. The scatter in the data is roughly one z-score unit.

TABLE 11: z-score comparisons.

Protein	ρ	slope	intercept	z_{opt}	z_{wt}
Adaptive walks with `NN` potential					
4icb	0.819	0.24 ± 0.005	3.20 ± 0.06	8.1	10.06
1ubq	0.896	0.41 ± 0.005	2.14 ± 0.05	8.7	9.22
1rro	0.932	0.31 ± 0.003	2.17 ± 0.04	8.8	10.88
1cew	0.938	0.25 ± 0.002	1.86 ± 0.03	7.9	6.20
2trxa	0.945	0.38 ± 0.003	2.13 ± 0.03	7.9	9.22
1lyz	0.919	0.37 ± 0.003	1.25 ± 0.04	7.5	7.70
Adaptive walks with `PROSA II` potential					
4icb	0.886	0.44 ± 0.007	-1.40 ± 0.07	6.6	8.08
1ubq	0.939	0.44 ± 0.003	-3.57 ± 0.06	5.1	6.40
1rro	0.938	0.49 ± 0.004	-1.16 ± 0.05	7.2	9.80
1cew	0.913	0.58 ± 0.006	-2.57 ± 0.06	6.9	5.91
2trxa	0.904	0.47 ± 0.006	-1.21 ± 0.06	7.3	8.06
1lyz	0.836	0.30 ± 0.005	0.71 ± 0.05	6.5	7.45

The sequences encountered along adaptive walks performed with one potential were evaluated with the other potential, see equation (6). z_{opt} denotes the best scores in the other potential, which should be compared to the corresponding wild type z-score z_{wt} and ρ is the correlation coefficient. The inaccuracies and inconsistencies of the two potentials are reflected by the fact that $a_{\mathtt{NN/PROSA}} \times a_{\mathtt{PROSA/NN}} \gg 1$ while, if the potentials were equivalent, we would observe that the product is exactly 1.

The data in Table 11 show that in some cases (`1cew`) we reach wildtype level in the reevaluated sequences that were produced using the other potential, while in

most cases the best scores are still one or two z-score units inferior. In general, the sequences produced with the NN potential score somewhat better in the PROSA than PROSA-optimized sequences do in NN. For example we almost reach PROSA-wildtype level for 2trxa and 1lyz with NN-optimized sequences, while the best PROSA-optimized sequences are still 0.8 z-score units short of the NN wildtype level.

7 Neutral Networks

The size of protein space makes it virtually impossible to check directly whether the neutral sets $S(\psi)$ form extensive connected networks, or whether they consist of a large number of disconnected isolated clusters. In previous studies we have introduced *neutral paths* as a tool to measure the connectedness of neutral sets [53, 3].

TABLE 12: Neutral Walks

Protein	n	$\mathcal{L}$		$\mathcal{L}/n$	
		NN	PROSA	NN	PROSA
1cbn†	46	36.1	44.6	0.786	0.956
1ubq†	76	64.7	72.5	0.851	0.954
1adr	76	−−	74.9	−−	0.986
4icb	76	68.2	73.6	0.898	0.968
1rro	108	95.9	105.4	0.879	0.976
2trxa†	108	87.5	106.3	0.810	0.984
1cew	108	100.8	106.7	0.898	0.988
1lyz†	129	115.4	126.2	0.894	0.978

Neutral Walks (see section 4.2) were performed with all the proteins dicussed in Chapter 5 except 1PGB and 1ROP to examine the extent of neutral paths in Sequence Space. Since the NN Potentials had problems integrating 1ADR into the fingerprint data base, there is no NN neutral walk data available for this structure (see Table 12 and Table 13. For each structure 10 neutral walks were performed. Table 12 shows the length of the neutral walks for both the PROSAand the NN data. PROSA data for the structures marked with † Table 12 was taken in from [3]. Table 13 shows the average distances of inverse folded sequences (10 neutral walks and 5 adaptive walks for each structure), $\langle d \rangle_{\mathrm{adw}}$, and the average distances between the endpoints of independent neutral walks. These distances ($\langle d \rangle_{\mathrm{nw}}$), are comparable to the sequence length in both potentials. Hence, we can conclude that neutral paths form connected neutral networks that extend through sequence space.

TABLE 13: Characteristics of Neutral Sets.

Protein	n	$\langle d \rangle_{\mathrm{adw}}/n$		$\langle d \rangle_{\mathrm{nw}}/n$	
		NN	PROSA	NN	PROSA
1cbn	46	−−	0.841	0.785	0.919
1ubq	76	0.896	0.803	0.851	0.872
1adr	76	−−	0.718	−−	0.904
4icb	76	0.539	0.731	0.898	0.854
2trxa	108	0.895	0.812	0.810	0.903
1rro	108	0.590	0.783	0.888	0.880
1cew	108	0.510	0.762	0.933	0.913
1lyz	129	0.916	0.822	0.895	0.920

Figures 49 to 56 show the distribution of z-scores encountered along neutral walks for four representative protein structures (`1rro`, `1cew`, `1ubq` and `4icb`) for both potential functions.

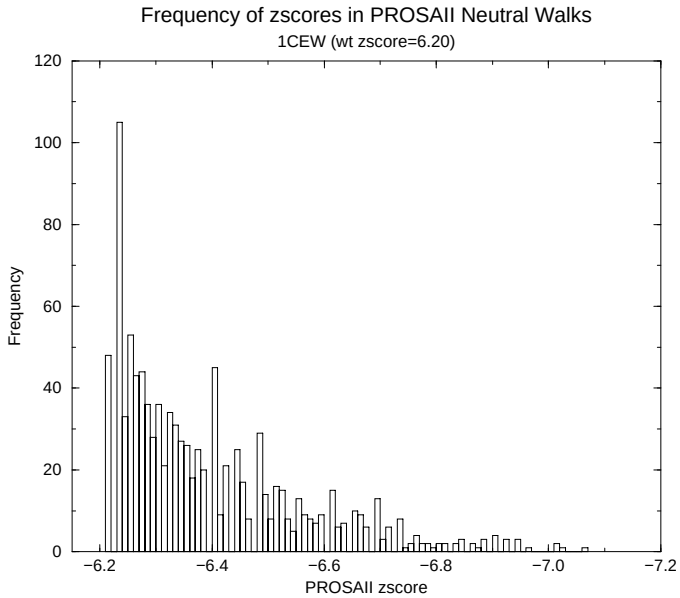

FIGURE 49: `1CEW`, distribution of z-scores along a neutral walk with the `PROSA II` potential

While the acceptable z-scores were restricted to a defined interval around the wild type z-score during the neutral walks performed with the `NN` potential, we implemented no explicit upper or lower bound on the `PROSA` z-scores in our neutral

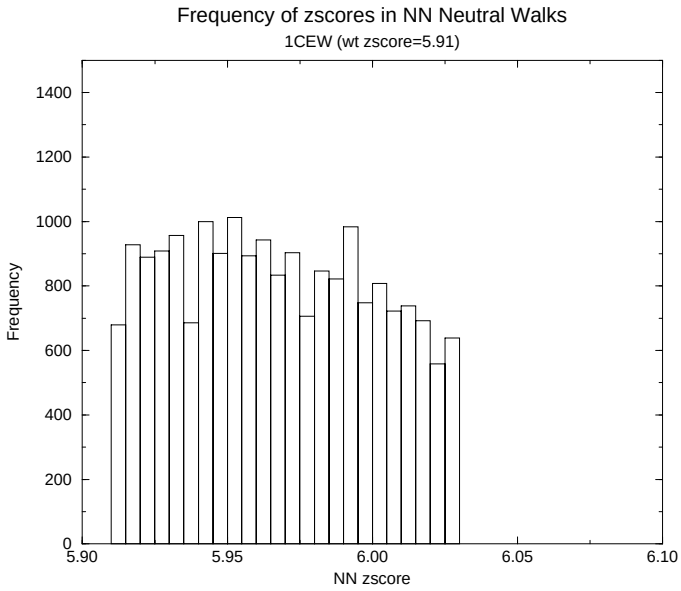

FIGURE 50: `1CEW`, distribution of z-scores along a neutral walk with the `NN` potential

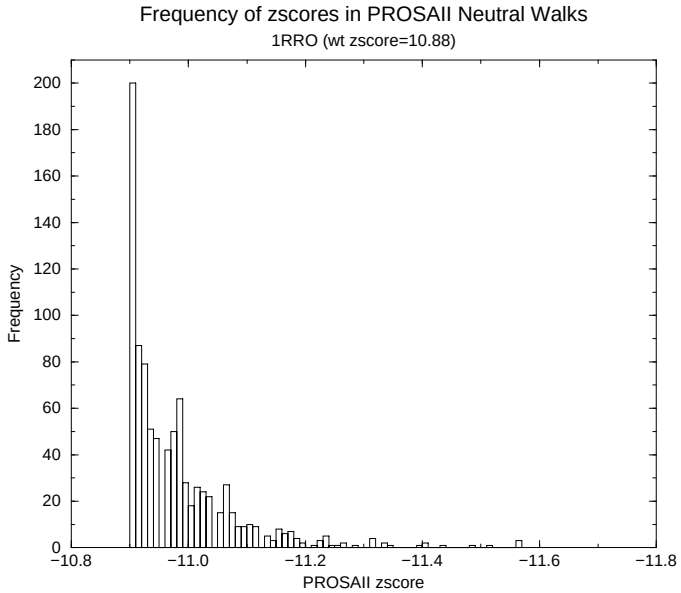

FIGURE 51: `1RRO`, distribution of z-scores along a neutral walk with the `PROSA II` potential

walk algorithms. Still, we find that the neutral walks are confined between wild-type level and only 1 z-score unit better than wildtype level.

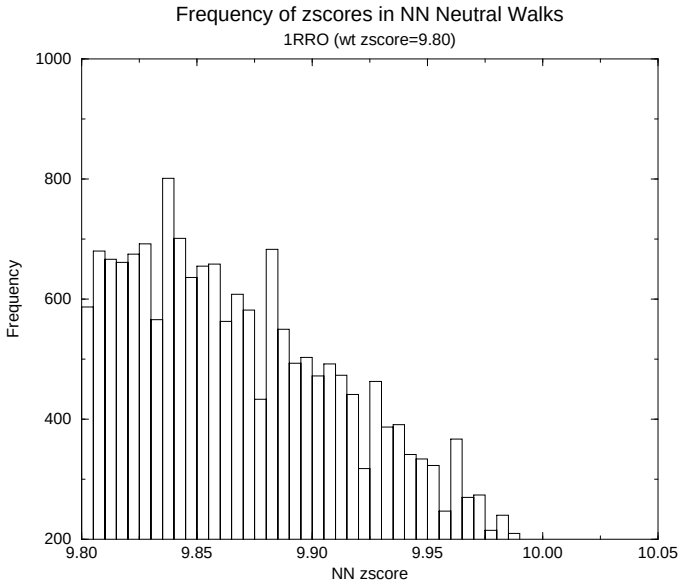

FIGURE 52: `1RRO`, distribution of z-scores along a neutral walk with the `NN` potential

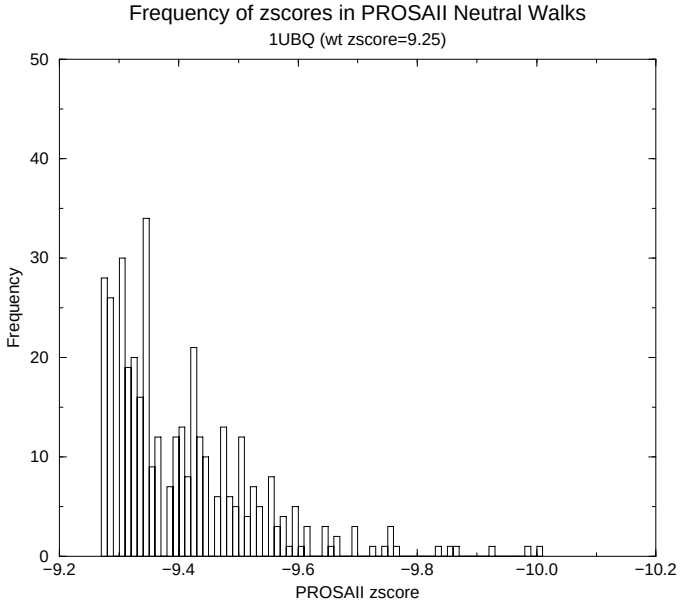

FIGURE 53: `1UBQ`, distribution of z-scores along a neutral walk with the `PROSA II` potential

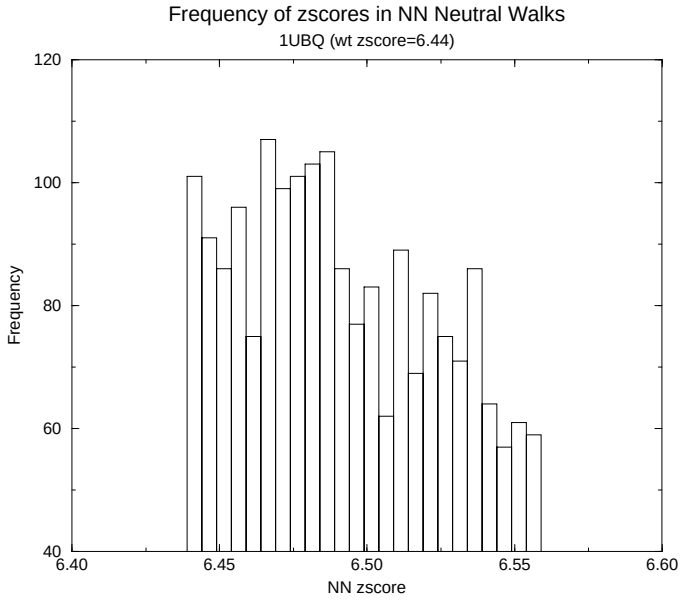

FIGURE 54: `1UBQ`, distribution of z-scores along a neutral walk with the `NN` potential

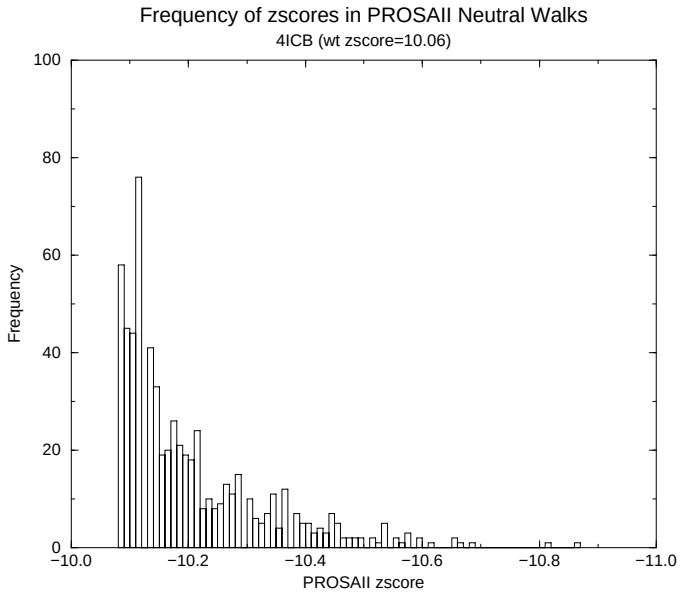

FIGURE 55: `4ICB`, distribution of z-scores along a neutral walk with the `PROSA II` potential

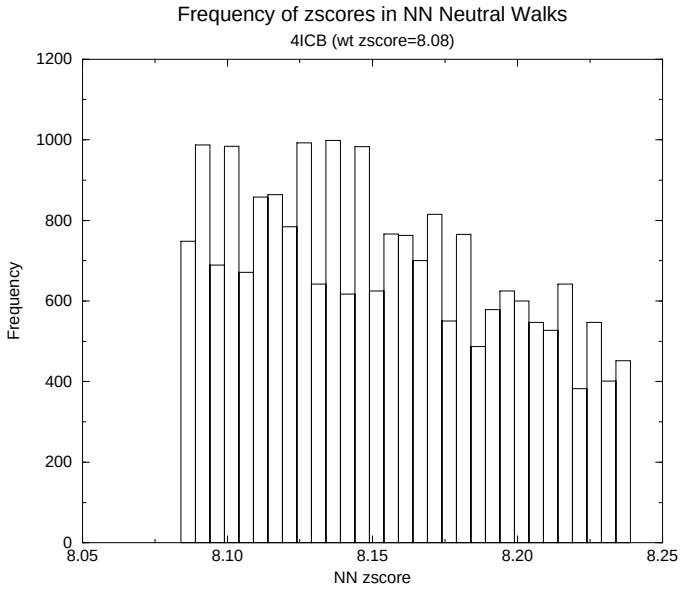

FIGURE 56: `4ICB`, distribution of z-scores along a neutral walk with the `NN` potential

8 Closest Approach and Shape Space Covering

A sequence-structure map exhibits *shape space covering* if it is possible to find almost all relevant folds within a small radius around almost any randomly chosen reference sequence. Similarly, we want to determine how close the neutral sets of two different native strucures $S(\psi)$ and $S(\varphi)$ come together. This is of particular interest if ψ and φ of two unrelated shapes, one, say containing only β-sheets and the other consisting of helices only. Due to the size of protein space, and of the neutral sets, we cannot determine the *closest approach distance*

$$D_f(S(\psi), S(\varphi)) = \min\{d(x, x')|x \in S(\psi) \text{ and } x' \in S(\varphi)\} \tag{7}$$

exactly.

However, consider a pair of walks $\{x_t\}$ and $\{y_t\}$ with the following properties:

(i) $x_0 \in S(\psi)$ and $y_0 \in S(\varphi)$.

(ii) For all t: $x_{t+1} \in S(\psi)$ is a neighbor of x_t that is closer to y_t, i.e., $d(x_{t+1}, y_t) \leq d(x_t, y_t)$ and analogously, $y_{t+1} \in S(\psi)$, $d(y_t, y_{t+1} = 1$, and $d(y_{t+1}, x_t) \leq d(y_t, x_t)$.

(iii) The equalities $d(x_{t+1}, y_t) = d(x_t, y_t)$ or $d(y_{t+1}, x_t) = d(t_t, x_t)$ hold at most M time-steps in a row, where M is a fixed constant.

The procedure terminates when no mutant $x_{n+1} \in S(\psi)$ of x_n can be found that is closer to y_n than x_n, and no mutant $y_{n+1} \in S(\varphi)$ can be found that is closer to x_n than y_n.

The residual distance $d(x_n, y_n)$, provides a (good) upper bound on the closest approach distance $D_f(\psi, \varphi)$ of the two networks. For the purpose of this section we have adopted a more restrictive definition of the neutral set of a shape ψ by requiring that the z-score is contained in a narrow interval:

$$S(\psi) = \{x \in \mathcal{Q}_{20}^n | 1.02z^* \geq z(x, \psi) \geq z^*\} \tag{8}$$

TABLE 14: Closest approach walks

Potential	D_f	D_f/n	D_1/n	D_2/n	$\overline{D_f}/n$	N
1cew/2trxa						
PROSA II	4.9	0.045	0.665	0.699	0.648	20
NN	5.2	0.048	0.838	0.843	0.785	10
2trxa/1rro						
PROSA II	6.5	0.060	0.755	0.711	0.711	13
NN	6.5	0.060	0.858	0.834	0.801	10

D_f: Hamming distance between the pairs of final sequences for each run, D_1 and D_2: Hamming distance of final sequences to their wildtype sequence, D_f/n, D_1/n, D_2/n: distances normalized by the sequence length n, $\overline{D_f}/n$ is the respective average and N the number of performed experiments

We start with a pair of sequences $x_0 \in S(\psi)$ and $y_0 \in S(\varphi)$. We use the wildtype sequence for both starting sequences x_0 and y_0. We attempt to find a mutant $x_1 \in S(\psi)$ of x_0 that is closer to y_0, and then mutant $y_1 \in S(\psi)$ of y_0 that is closer to x_1. The procedure is repeated until no mutant $x_{n+1} \in S(\psi)$ of x_n can be found that is closer to y_n than x_n, and no mutant $y_{n+1} \in S(\varphi)$ can be found that is closer to x_n than y_n. In order to increase the efficiency of the simulations we allow these *closest approach walks* to accept a fixed number of mutants that lie within $S(\psi)$ and at least do not increase the Hamming distance to y_0 before terminating the walk.

From our simulations we derive the following quantities:

(i) D_f is the Hamming distance between the pairs of final sequences for each run;

(ii) D_1 and D_2 is the Hamming distance of the two final sequences to their respective native (wildtype) sequence;

(iii) d_f is the average Hamming distance between the final sequences from all different runs.

The values of $D_{1,2}$ and d_f complement the information about the extent of the neutral networks, verifying that they are indeed spanning most of sequence space, see Table 12.

While small closest approach distances $D_f(\psi, \varphi)$ even for vastly different structures ψ and φ are a necessary condition for shape space covering, they cannot serve as a proof. For a particular reference sequence x let $d(x, \psi)$ be the minimum Hamming distance between x and a sequence that fold into ψ. We define the *covering radius* R_c as the average of $d(x, \psi)$ over all sequences x and structures ψ. Consequently, we expect to find most folds within a ball of radius R_c centered at a typical point in sequence space.

TABLE 15: Shape Space Covering

Protein	D_y	D_y/n	D_x	D_x/n	$\overline{D_{xr}}$	$\overline{D_{xr}}/n$	N
`1ubq`	61.44	0.808	62.67	0.825	34.80	0.457	10
`4icb`	63.22	0.831	61.89	0.814	36.60	0.481	10
`1cew`	90.21	0.835	95.88	0.888	36.80	0.341	10
`2trxa`	93.44	0.865	94.44	0.874	50.10	0.463	10
`1rro`	89.11	0.825	90.22	0.854	55.40	0.513	10
`1lyz`	113.78	0.882	114.78	0.890	47.90	0.371	10

D_y: Hamming distance of final sequences of adaptive walks among each other, D_x: Hamming distance of final sequences of closest approach walks among each other, $\overline{D_{xr}}$: Hamming distance to random sequences (starting points of adaptive walks), D_y/n, D_x/n, $\overline{D_{xr}}/n$: Hamming distances normalized by sequence length n; N: number of performed experiments

Upper bounds for $d(x, \psi)$ can be readily obtained in a variation of the closest approach walks: A sequence y folding into ψ is constructed by an adaptive walk (initialized at x). Starting from y we then perform a neutral walk with target x and measure the residual distance. Obviously this procedure is computationally quite demanding as we need to perform the calculation for a large number of structures and reference sequences. In the case of RNA we found that the covering radius is surprisingly small and dominated by the logic of the base pairing rules

[53, 31]. Since proteins have no equivalent to the restrictive RNA base pairing rules, we conjecture that the covering radius for proteins will be at least as small.

A number of prliminary computations were performed for 6 different proteins (`1cew`, `1lyz`, `2trxa`, `1rro`, `1ubq`, `4icb`). The procedure was the following: Adaptive walks were conducted starting from a random sequence and terminated at a treshold z-score (z^*) ca. 3 units better than the wildtype z-score (z_{wt}). A closest approach run with the final sequences y of these runs was then performed in the direction of the first random sequence r of the adaptive walk to receive an estimate for the distance D_{xr} which is the distance from a random point in sequence space to a sequence with native like z-score. Table 15 shows the results for 10 runs for each protein with the `PROSA II` potential.

The average $\overline{d_\psi}$ of $d(x,\psi)$ over the reference sequences can be used to estimate the size of the neutral set $S(\psi)$, since a ball of radius $\overline{d_\psi}$ contains on the order of 1 sequence folding into ψ. Thus,

$$|S(\psi)| \approx \frac{20^n}{B(\overline{d_\psi})}, \tag{9}$$

where $B(r) \approx 19^r \binom{n}{r}$ is the volume of a ball of radius r in sequence space. The number of sequences folding into ψ has been used as a measure of *designability* in the context of lattice proteins [32, 39] hence the designability decreases with $\overline{d_\psi}$.

9 Janus

In a quite spectacular experiment Dalal *et al.* [17] have designed a protein, `Janus` that has 50% sequence identity with a predominantly β-sheet protein (`1pgb`), but which adopts a four-helix bundle conformation (the structure of the dimer of `1rop`) and possesses the attributes of a native protein. Starting from the `1pgb` sequence they mutated selected residues to adapt the sequence to the Rop fold, usually by replacing them with the corresponding amino acid from Rop (only 9 positions in Janus differ from both Rop and `1pgb`).

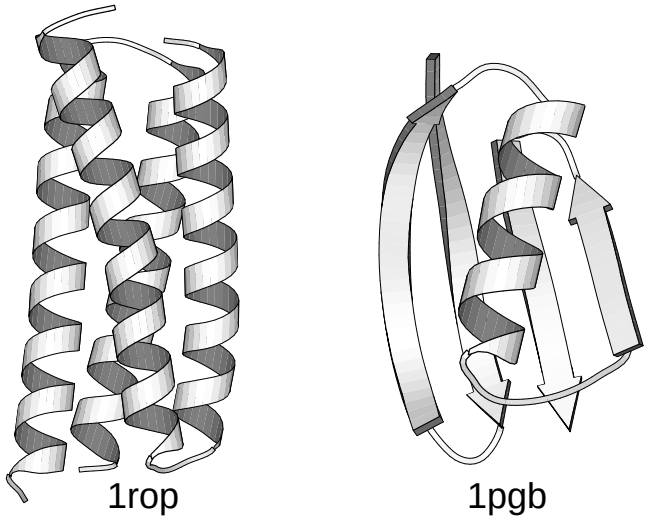

FIGURE 57: Schematic drawing of the structures of `1ROP` and `1PGB`

This is an excellent experimental example which addresses the questions concerning sequence structure relationships we discuss throughout this work. We have therefore designed a computer experiment that mimics the Dalal *et al.* procedure. Our approach is to define a mutated Rop sequence which has high sequence identity to the `1pgb` sequence, but which retains a Rop wild type z-score when evaluated on the Rop structure. In analogy to the experimental procedure we restrict the sequence to be identical to either Rop or `1pgb` in each position. The 7 amino acids at the C-terminus of Rop are ignored. The native form of Rop is a dimer. In an isolated monomer many amino acids buried in the dimer would be exposed, giving rise to non-sensical z-scores. Since our empirical potentials are

not equipped to deal with dimers, we have artificially connected the two chains of 1rop resulting in a good z-score for the wild-type sequence (z-score of the monomer -4.94, z-score of the dimer -7.47).

TABLE 16: z-score evaluation of wild type and Janus sequences with the PROSA II potential

	Structure	
Sequence	1rop	1pgb
1pgb	-1.61	7.62
Janus (Dalal)	6.17	2.36
Janus (simulation)	7.52	2.00
1rop	7.47	0.72

In our computer experiment, both copies of the connected sequence forming the Rop dimer are always mutated together. An attempted mutation in the 1rop dimer is defined by replacing an amino-acid with the corresponding amino-acid from 1pgb. The mutation is accepted if the z-score of the mutated sequence is no worse than the wild type z-score of 1rop. When using the PROSA potential we require that both C^{α}- and C^{β}-scores are as good as the wildtype. The experiment is terminated when no further acceptable mutations can be introduced.

We performed 18 such simulations, some of the resulting sequence are shown in figure 59. The sequences have an average Hamming distance of 25.1 to the wildtype 1pgb sequence, amounting to a percentage sequence identity of 55.2%. This is only slightly larger than the experimental value: Janus was designed to have 50% sequence identity with 1rop. Our results indicate that Dalal *et al.* indeed employed a near minimal number of mutations from 1pgb.

One should note, however, that sequences much closer to 1pgb could be found without the restriction that sequences must be identical to either 1rop or 1pgb in each position, in which case hamming distances of 13 or 14 were obtained. The average Hamming distance to the wildtype 1rop sequence is 26.9, amounting to 52.0% sequence identity, which is also quite similar to value for the Janus sequence of 41%. Instead of terminating the runs when no mutations can be introduced that lead closer to 1pgb, we may allow other mutations that yield a native-like

score for 1rop. This parallels the procedure for the closest approach walk.

```
MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARF[GDDGENL]  (1rop)
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE          (1pgb)
MTKKAILALNTAKFLRTQAAVLAAKLEKLGAQEANDNAVDLEDTADDLYKTLLVLA          (Janus)
```

FIGURE 58: Wildtype sequences of 1ROP and 1PGB, as well as the Janus sequence

Figure 59 shows nine of the 18 Prosa II-evolved janus sequences. In the consensus line below '*', ':', and '.' mark conserved or nearly conserved positions.

Several sequence positions were conserved in all or most of our simulations and also agreed with the amino acids chosen by Dalal *et al.* [17]. Among the conserved amino-acids was the Asp 46 residue found in 1ROP, which participates in the intra-monomer salt bridge with Arg 16, however the Arg 16 residue was replaced by Thr in 10 of the 18 sequences.

TABLE 17: z-score evaluation and comparism of wild type and Janus se quen ces with PROSA II and the Vienna Tesselation potentials

	Structure			
	PROSA		Tessellation	
Sequence	1rop	1pgb	1rop	1pgb
1pgb	-1.61	7.62	0.07	3.47
Janus (Dalal)	6.17	2.36	7.10	2.90
Janus (simulation)	7.52	2.00	5.49	2.66
1rop	7.47	0.72	5.31	0.47

Similar computer simulations were performed with the Vienna Tesselation Potential developed by G. Weberndorfer [63, 64] at our institute, to test whether our results were reproducable in another potential. The procedure of this simulations was identical to those with the PROSA II potential. We found that sequences with more than 50% sequence identity to 1pgb could be generated with the Tesselation Potential. The 10 final sequences from independent experiments showed an average sequence identity of approximately 87% while still retaining wild-type like z-scores on the 1rop structure [64]. The z-scores of the wild-type and the

```
MTKKEILNLKMLKGIRTTTATLAETLNKLDKDYANDIGEDLEDHYDDLYKTFTATF
MTKKEILALKTLKGITTQTATLAETLNKLDKDYANDNCEDLHDTADDLTKTFTATF
MTYKEKTALKTLKGITTTTAVLAETLNKVDKDYANDICEDLEDHADDLYKTFTATE
MTKKEKLNLKTLKGITTTTAVLAETLNKVDADYANDICEDLEWHADDLTKTFTATF
MTYKEITALKTLKGIRTTTAVLAETLNKVDKDYANDICEDLHDHADEATKTFTATF
MTYKEITALKMAKGITTQTAVLAETLEKVDKDYANDIGEDLEWTADELYKTFTATF
MTKKEITALKMLKGITTTTAVLLETLNELDADYAADICEDGEWHYDELTKTFTATF
MTKKEILALKTAKGIRTTTAVLLEKANKVDKDYAADIGVDLHDHADDLYKTFTATF
MTYKEKTALKMAKFIRSTTLVLLEKANKLDADYQADICEDLHWHADELYRTFTARF
** **   **  * * : * .* *. :::* **  *   * .   *:  :****
```

FIGURE 59: PROSA-evolved Janus sequences.

average for the 10 Janus sequences calculated with both PROSA II and the Tesselation Potential can be found in Table 17. The sign of the PROSA II z-scores was reversed for better readability, better scores are more poitive in this Table.

10 Conclusion and Outlook

Although some differences appear in detail, the behavior of adaptive walks, neutral walks, and closest approach walks, and consequent implications such as the existence of extensive neutral networks and shape space covering, are common to both the PROSA II and the neural network NN potentials. Hence our conclusions concerning the topology of sequence space, as defined by the various types of walks, are independent of the details of any one potential.

We have found a comparably good correlation between the NN and PROSA potential [2] and preliminary experiments also showed a good correlation between the tessellation potential and PROSA [64] However, usually sequences optimized in one potential exhibit significantly worse z-scores in another. This reflects the inaccuracies and inconsistencies between different potentials. Sequences should therefore be optimized until a z-score several units better than the wild type level is reached. Zhang and Skolnick [67] have estimated that the z-score of a protein of 100 residues should be better than 15, significantly larger than the score in currently available potentials.

We found that neutral paths within the sets $S(\psi)$ extend to almost the length of the amino acid sequence. We therefore conclude that neutral sets form extensive *neutral networks* that percolate the entire sequence space. The existence of extensive neutral networks meets a claim raised by Maynard-Smith [42] for protein spaces that are suitable for efficient evolution. Empirical evidence for a large degree of *functional* neutrality in protein space is indeed observed [41]. The existence of extensive neutral networks has been established using both the PROSA II and the NN potential. Nevertheless it will be necessary to produce neutral path data from the tessellation potential for comparison. In addition, the length distribution of neutral paths can be compared with simple random landscapes models [48, 50] in order to detect systematic deviations that hint at anisotropies in protein space.

It will be necessary to further study the dependence of neutral walk lengths on z-score for different refinements of the potential. Neutral walk experiments should be performed at z-score levels up to at least 3 units better than the wild-type in

order to obtain more detailed information on the diameter of neutral networks.

A more direct, but also more demanding, way of determining whether $S(\psi)$ is indeed connected, as predicted by random graph models [48], is the following: Two independent inverse folding runs are performed for the same structure to produce two distant members of $S(\psi)$. We then try to explicitly construct a neutral path connecting the two sequences. This can be done by a variant of the closest approach procedure described in chapter 8. The feasibility of such an approach has been demonstrated in the RNA case [25].

Like the existence of neutral networks, shape space covering was first observed in computational studies of the RNA sequence structure relationship [53, 30, 31]: any native structure can be found within a small ball in sequence space that can be centered at an arbitrary reference point. Sander and Schneider [51] have argued that sequences with more than some 30% sequence homology will give rise to the same fold. On the other hand, the `Janus` examples shows that exceptions to this rule can be constructed. Our computational data support an even stronger claim: sequences that fold into two completely different native structures need not differ by more than a few crucial amino acids. So far, we have only a very limited sample of closest approach data due to the high computational costs of both the `PROSA` and the `NN` potential. More extensive studies are feasible with G. Weberndorfers [63, 64] implementation of the tessellation potential. With this potential a much larger set of structure pairs should be analyzed. Furthermore, closest approach walks yield an upper bound on $D_f(\psi, \varphi)$. More extensive calculations (in particular walks with a large number M of steps that do not decrease Hamming distance) will be necessary to obtain improved bounds. In addition closest approach experiments should be performed at different z-score levels in order to obtain more reliable estimates for the closest approach distance $D_f(\psi, \varphi)$, analogous to the neutral walk data shown in Table 12.

The random graph theory of neutral networks predicts $D_f(\psi, \varphi) = 1$, i.e. the neutral networks of any two structures ψ and ϕ touch each other [48]. Although we found values larger that 1 for $D_f(\psi, \varphi)$ in our closest approach experiments, $D_f(\psi, \varphi)$ was surprisingly small. From our results we can conclude that neutral networks in protein space come very close together and further experiments

should be performed to consolidate these findings. It will be interesting to see if there is a dependence of D_f on measure dissimilarity between the structures ψ and ϕ. While small closest approach distances $D_f(\psi, \varphi)$ even for vastly different structures ψ and φ are a necessary condition for shape space covering, they cannot serve as a proof. For a particular reference sequence x let $d(x, \psi)$ be the minimum Hamming distance between x and a sequence that fold into ψ. We define the *covering radius* R_c as the average of $d(x, \psi)$ over all sequences x and structures ψ. Consequently, we expect to find most folds within a ball of radius R_c centered at a typical point in sequence space.

Upper bounds for $d(x, \psi)$ can be readily obtained in a variation of the closest approach walks: A sequence y folding into ψ is constructed by an adaptive walk (initialized at x). Starting from y we then perform a neutral walk with target x and measure the residual distance. Obviously this procedure is computationally quite demanding as we need to perform the calculation for a large number of structures and reference sequences. In the case of RNA we found that the covering radius is surprisingly small and dominated by the logic of the base pairing rules [53, 31]. In the preliminary experiments we performed, we found $d(x, \psi)$ to be small but larger than in the RNA case, it will be necessary to perform more experiments of this kind to establish a tighter upper bound.

The question whether designability is an intrinsic property of a (native) fold ψ should be investigated further by trying to construct sequences with native-like z-scores from restricted alphabets. Only if the ordering of a sample of structures with respect to the lengths of adaptive walks (and other measure of designability, such as $|S(\psi)|$) is the same for different alphabets, designability can considered as a well-defined property of a structure. Preliminary data suggest that this is indeed the case [3], these data also indicates that one can distinguish between alphabets that allow the design of (most) native folds and alphabets that cannot be used to build native-like protein structures. This question is of particular interest in the context of the origin of life and evolution of the genetic code [21, sect. XIV.4]. While native-like proteins can be designed from reduced alphabets, recent experiments [18, 47] as well as computer simulations [3, 11] suggest that two letters are not sufficient. Extensive studies of adaptive walks, neutral walks,

and closest approach experiments with restricted alphabets should be performed. It will be necessary to consider a large number of different amino acids alphabets to arrive at a conclusive answer.

The evolutionary implications of neutral networks and shape space covering are discussed in detail in [37, 35]. Extensive neutral network set the stage for an efficient exploration of sequence by diffusion on the neutral network. Shape space covering, on the other hand, guarantees a constant rate of exploring novel structures that have not been encountered before. The rate of exploration begins to slow down only when a sizable set of all shapes have already been realized.

11 List of Figures

1 Tessellations in 2 dimensions 17

2 Comparism of adaptive walks with the NN potential and the PROSA II potential. 26

3 Adaptive walks with the NN potential (solid lines) and the PROSA II potential (dotted lines). 26

4 DNA-binding domain of P22, PDB Structure 1ADR 28

5 Ubiquitin from Human Erythrocytes, PDB Structure 1UBQ 29

6 Calbindin from Bos Taurus, PDB Structure 4ICB 31

7 Thioredoxin from E. Coli, PDB Structure 2TRXA 32

8 Cystatin from hen egg white, PDB Structure 1CEW 33

9 Oncomodulin from rat tumours, PDB Structure 1RRO 34

10 Lysozyme from chicken egg white, PDB Structure 1LYZ 35

11 PDB Structure 1ROP and 1PGB 36

12 1ADR, 10 Adaptive Walks with PROSA II 39

13 1UBQ, 10 Adaptive Walks with PROSA II 40

14 1UBQ, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 40

15 Regression of 10 Adaptive Walks in Figure 14 41

16 1UBQ, 5 Adaptive Walks with the NN Potential 41

17 1UBQ, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 42

18 Regression of 5 Adaptive Walks in Figure 17 43

19 4ICB, 10 Adaptive Walks with PROSA II 43

20 4ICB, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 44

21 Regression of 10 Adaptive Walks in Figure 20 45

22 4ICB, 5 Adaptive Walks with the NN Potential 45

23 4ICB, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 46

24 Regression of 5 Adaptive Walks in Figure 23 46

25 1CEW, 10 Adaptive Walks with PROSA II 47

26 1CEW, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 48

27 Regression of 10 Adaptive Walks in Figure 26 48

28 1CEW, 5 Adaptive Walks with the NN Potential 49

29 1CEW, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 50

30 Regression of 5 Adaptive Walks in Figure 29 50

31 1RRO, 10 Adaptive Walks with PROSA II 51

32 1RRO, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 52

33 Regression of 10 Adaptive Walks in Figure 32 52

34 1RRO, 5 Adaptive Walks with the NN Potential 53

35 1RRO, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 53

36 Regression of 5 Adaptive Walks in Figure 35 54

37 2TRXA, 10 Adaptive Walks with PROSA II 55

38 2TRXA, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 55

39 Regression of 10 Adaptive Walks in Figure 38 56

40 2TRXA, 5 Adaptive Walks with the NN Potential 57

41 2TRXA, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 57

42 Regression of 5 Adaptive Walks in Figure 41 58

43 1LYZ, 10 Adaptive Walks with PROSA II 58

44 1LYZ, 10 Adaptive Walks with PROSA II evaluated with the NN Potential 59

45 Regression of 10 Adaptive Walks in Figure 44 59

46 1LYZ, 5 Adaptive Walks with the NN Potential 60

47 1LYZ, 5 Adaptive Walks with the NN Potential evaluated with PROSA II 61

48 Regression of 5 Adaptive Walks in Figure 47 61

49 1CEW, distribution of z-scores along a neutral walk with the PROSA II potential 65

50 1CEW, distribution of z-scores along a neutral walk with the NN potential 66

51 1RRO, distribution of z-scores along a neutral walk with the PROSA II potential 66

52 1RRO, distribution of z-scores along a neutral walk with the NN potential 67

53 1UBQ, distribution of z-scores along a neutral walk with the PROSA II potential 67

54 1UBQ, distribution of z-scores along a neutral walk with the NN potential 68

55 4ICB, distribution of z-scores along a neutral walk with the PROSA II potential 68

56 4ICB, distribution of z-scores along a neutral walk with the NN potential 69

57 Schematic drawing of the structures of 1ROP and 1PGB 74

58 Wildtype sequences of 1ROP and 1PGB, as well as the Janus sequence 76

59 PROSA-evolved Janus sequences. 77

12 List of Tables

1 Average length ℓ of Adaptive Walks to reach wildtype z-score. . . 27

2 Structure Motifs of `1ADR` 29

3 Structure Motifs of `1UBQ` 30

4 Structure Motifs of `4ICB` 31

5 Structure Motifs of `2TRXA` 33

6 Structure Motifs of `1CEW` 34

7 Structure Motifs of `1RRO` 35

8 Structure Motifs of `1Lyz` 36

9 Structure Motifs of `1ROP` 37

10 Structure Motifs of `1PGB` 37

11 z-score comparisons. 62

12 Neutral Walks 64

13 Characteristics of Neutral Sets. 65

14 Closest approach walks 71

15 Shape Space Covering 72

16 z-score evaluation of wild type and `Janus` sequences with the `PROSA II` potential 75

17 z-score evaluation and comparism of wild type and `Janus` se quen ces with `PROSA II` and the Vienna Tesselation potentials 76

13 References

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[2] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler. Exploring protein sequence space using knowledge based potentials. *Protein Science*, 1998. submitted, Santa Fe Institute preprint 98-11-103.

[3] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 1997.

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

[5] R. T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London*, 53:370–418, 1963.

[6] W. Bode and V. Turk. The cystatins: Protein inhibitors of cysteine proteinases. *FEBS Lett.*, 285:213–219, 1991.

[7] 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*, 7:257–264, 1990.

[8] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–169, 1991.

[9] C. Bradford, Barber, D. P. Dobkin, and H. T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22:469–421, 1996.
`URL: http://www.acm.org`.

[10] S. Bryant and C. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 16:92–112, 1993.

[11] N. E. G. Buchler and R. A. Goldstein. The effect of alphabet size and foldability requirements on protein structure designability. *Proteins*, 1999. in press.

[12] 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.

[13] H. S. Chan and K. A. Dill. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.*, 92,5:3118–3135, 1990.

[14] H. S. Chan and K. A. Dill. Comparing folding codes for proteins and polymers. *Proteins*, 24:335–344, 1996.

[15] C. Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[16] G. M. Crippen and Y. Z. Ohkubo. Statistical mechanics of protein folding by exhaustive enumeration. *Proteins*, 32:425–437, 1998.

[17] S. Dalal, S. Balasubramanian, and L. Regan. Protein alchemy: Changing β-sheet into α-helix. *Nat. Struct. Biol.*, 4(7):548–552, 1997.

[18] A. R. Davidson, K. J. Lumb, and R. T. Sauer. Cooperatively folded proteins in random sequence libraries. *Nat. Struc. Biol.*, 2:856–863, 1995.

[19] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yeo, P. D. Thomas, and H. S. Chan. Principles of protein folding: a perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[20] K. E. Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. (USA)*, 78:5275–5278, 1981.

[21] M. Eigen and P. Schuster. *The Hypercycle – A Principle of Natural Self-Organization.* Springer-Verlag, Berlin, 1979.

[22] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatsh.Chem.*, 122:795–819, 1991.

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

[24] 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 landscapes and combinatory landscapes. *Phys.Rev.E*, 47:2083–2099, 1993.

[25] U. Göbel, C. V. Forst, and P. Schuster. Structural constraints and neutrality in RNA. In R. Hofestädt, T. Lengauer, and D. S. M. Löffler, editors, *Proceedings of the German Conference on Bioinformatics 1996*, volume 1278 of *Lecture Notes in Computer Science*, pages 156–165, Berlin, Heidelberg, New York, 1997. Springer Verlag.

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

[27] 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.

[28] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Optimized energy functions for tertiary structure prediction and recognition. In H.Bohr and S.Brunak, editors, *Protein Structure by Distance Analysis*. IOS Press, 1994.

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

[30] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence struc-

ture maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996. SFI preprint 95-10-099.

[31] 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.*, 127:375–389, 1996. SFI preprint 95-10-099.

[32] C. T. Hao Li, Robert Helling and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.

[33] 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.

[34] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, 25:231–234, 1997.

[35] M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

[36] M. A. Huynen and P. Hogeweg. Pattern generation in molecular evolution. Exploitation of the variation in RNA landscapes. *J. Mol. Evol.*, 39:71–79, 1994.

[37] 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. SFI preprint 95-01-006, LAUR-94-3763.

[38] S. A. Kauffman. *The Origin of Order*. Oxford University Press, New York, Oxford, 1993.

[39] H. Li, C. Tang, and N. S. Wingreen. Are protein folds atypical? *Proc. Natl. Acad. Sci. (USA)*, 95:4987–4990, 1998.

[40] R. Luthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[41] M. A. Martinez, V. Pezo, P. Marlière, and S. Wain-Hobson. Exploring the functional robustness of an enzymeb y *in vitro* evolution. *EMBO J.*, 15:1203–1210, 1996.

[42] J. Maynard-Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[43] S. Miyazawa and R. L. Jernigan. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[44] P. J. Munson and R. K. Singh. Statistical significance of hierarchical multi-body potentials based on delauney tessellation and their application in sequence-structure alignment. *Protein Science*, 6:1467–1481, 1997.

[45] A. Neumaier, S. Dallwig, W. Huyer, and H. Schichl. New techniques for the construction of residue potentials for protein folding. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes Comput. Sci. Eng.* Springer, Berlin, 1999.

[46] J. Novotny, R. Bruccoleri, and M. Karplus. An analysis of incorrectly folded protein models. implications for structure predictions. *J. Mol. Biol.*, 1984.

[47] K. Plaxco, D. Riddle, V. Grantcharova, and D. Baker. Simplified proteins: minimalist solutions to the "protein folding problem". *Curr. Opin. Struct. Biol.*, 8:80–85, 1998.

[48] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997. SFI preprint 95-07-058.

[49] C. M. Reidys. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.*, 19:360–377, 1997.

[50] C. M. Reidys and P. F. Stadler. Neutrality in fitness landscapes. *Appl. Math. & Comput.*, 1998. submitted, Santa Fe Institute preprint 98-10-089.

[51] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[52] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *Journal of Biotechnology*, 41:239–257, 1995.

[53] 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. Lond. B*, 255:279–284, 1994.

[54] P. Schuster, P. F. Stadler, and A. Renner. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol.*, 7:229–235, 1997.

[55] R. K. Singh, A. Tropsha, and I. I. Vaisman. Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues. *J. Comput. Biol.*, 3:213–221, 1996.

[56] 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.

[57] 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.

[58] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[59] M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and F. Hannes. Helmholz free energies of atom pair interactions in proteins. *Folding & Design*, 1:289–298, 1996.

[60] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *Eur. Biophys. J.*, 23:29–38, 1994.

[61] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[62] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257:457–469, 1996.

[63] G. Weberndorfer. Empirical protein potentials from delauney tessellation. Master's thesis, Universtät Wien, 1999.

[64] G. Weberndorfer, I. L. Hofacker, and P. F. Stadler. An efficient potential for protein sequence design. In *GCB '99: German Conference on Bioinformatics*, 1999.

[65] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chemistry*, 7:230, 1986.

[66] M. Wilmanns and D. Eisenberg. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. *Proc. Natl. Acad. Sci. (USA)*, 90:1379–1383, 1993.

[67] L. Zhang and J. Skolnick. What should the z-score of native protein structures be? *Protein Science*, 7:1201–1207, 1998.

[68] W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. Statistical geometry analysis of proteins: implications for inverted structure prediction. In L. Hunter and T. Klein, editors, *Biocomputing: Proceedings of the 1996 Pacific Symposium*, pages 614–23. World Scientific Publishing Co, 1996.

[69] W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. A new approach to protein fold recognition based on delaunay tessellation of protein structure. In L. Hunter and T. Klein, editors, *Biocomputing: Proceedings of the 1997 Pacific Symposium*, pages 486–97. World Scientific Publishing Co, 1997.

CURRICULUM VITAE

Mag. Aderonke Babajide

geb. 03. 06. 1966, wohnhaft: Wien IX, Österreich

Österreichische Staatsbürgerin

Vater: Dr. Abayomi Diekola Babajide

Mutter: Marie-Louise Babajide

Volksschule für Knaben und Mädchen XVI Wien: 09. 1972 – 06. 1974
Deutsche Schule Lagos, Nigeria: 09. 1974 – 18. 06. 1985
Universität Wien: 1985 – 1999

- Diplomstudium Chemie Studienzweig Biochemie 01. 10. 1985 – 25. 04. 1996
- Mitglied der Fakultätsvertretung NAWI 1987 – 1989
- Diplomarbeit: *"Inverse Folding of Proteins with Knowledge Based Potentials of Mean Force"*, am Institut für Theoretische Chemie und Strahlenchemie bei Doz. Dr. Peter F. Stadler, 11. 1994 – 04. 1996
- 2. Diplomprüfung am 25. 04. 1996
- Sponsion zur Magistra rerum naturalium am 10. 10. 1996.
- Dissertation: *"Neutral Networks in Protein Space"*, am Institut für Theoretische Chemie und Molekulare Strukturbiologie bei A.o. Univ. Prof. Dr. Peter F. Stadler, 05. 1996 – 12. 1999.

Forschungsaufenthalte:

1) Workshop on Molecular Biotechnology, Jena, 28. 11. 1994 – 02. 12. 1994
2) The Santa Fe Institute, Santa Fe, New Mexico, 21. 05. 1996 – 21. 06. 1996
3) The Santa Fe Institute, Santa Fe, New Mexico, 28. 08. 1997 – 30. 09. 1997
4) School for Computational Biology, Lipari, Italien, 20. 06. 1999 – 04. 04. 1999