Exploring Protein Sequence Space Using Knowledge-Based Potentials

Aderonke Babajide
Robert Farber
Ivo L. Hofacker
Jeff Inman
Alan S. Lapedes

SFI WORKING PAPER: 1998-11-103

SFI Working Papers contain accounts of scientific work of the authors and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
Exploring Protein Sequence Space Using Knowledge Based Potentials

ADERONKE BABAJIDE\textsuperscript{a}, ROBERT FARBER\textsuperscript{c,d} IVO L. HOFACKER\textsuperscript{a}, JEFF INMAN\textsuperscript{b,d}, ALAN S. LAPIDES\textsuperscript{c,d}, AND PETER F. STADLER\textsuperscript{a,d}

\textsuperscript{a}Institut für Theoretische Chemie, Universität Wien  
Währingerstraße 17, A-1090 Wien, Austria  
Phone: **43 1 4277 52742  Fax: **43 1 4277 52793

\textsuperscript{b}National Center for Genome Resources  
1800-A Old Pecos Trail Santa Fe, NM 87505  
Phone: (505) 982-7840  Fax: (505) 995-4432

\textsuperscript{c}Theoretical Division, MS B213 LANL  
Los Alamos National Laboratory, T-13, Los Alamos NM 87545  
Phone: (505) 667 7608  Fax: (505) 665 3003

\textsuperscript{d}The Santa Fe Institute  
1399 Hyde Park Road, Santa Fe, NM 87501, USA  
Phone: (505) 984 8800  Fax: (505) 982 0565

*Address for correspondence:  
Peter F. Stadler, Institut f. Theoretische Chemie, Univ., Wien  
Währingerstraße 17, A-1090 Wien, Austria  
Phone: **43 1 4277 52737  Fax: **43 1 4277 52793

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, M. Sippl’s PROSA pair potential and a neural network based potential, are used to verify these claims.

Keywords: Knowledge-Based Potentials, Inverse Folding, Neutral Networks, Protein Evolution
1. Introduction

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 (Schuster, 1995; Huynen et al., 1996).

Protein space, on the other hand, is still largely *Terra incognita*. Considering merely the 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 (Chothia, 1992; Holm and Sander, 1997). Focusing on the global features of the sequence-structure map of polypeptides, the questions will be of a statistical nature, dealing with ensembles of sequences and structures rather than targeting a particular sequence or structure of a particular protein: How many sequences adopt the same fold? How are these sequences distributed in sequence space? How similar can sequences be that fold into very different shapes?

In a previous contribution (Babajide et al., 1997) we have argued that knowledge-based potentials can be used in principle to answer this type of questions. In fact, knowledge-based potentials are designed to recognize whether a sequence $x$ folds into a native structure $\psi$. 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 (Drexler, 1981; Bowie et al., 1991).

In this contributions we extend earlier results in two directions: (i) Repeating earlier computer experiments using a very different form of the potential function we arrive at the same qualitative results. (ii) We answer the main question that was left open in (Babajide et al., 1997), namely whether there is *shape space covering*, that is whether the neutral networks of any two different shapes come close to each other.

2. Knowledge Based Potentials

Recent studies using knowledge based potentials (Bauer and Beyer, 1994; Bowie et al., 1991; Godzik et al., 1992; Goldstein et al., 1992; Grossman et al., 1995; Henelich et al., 1990; Sippl, 1993a,b) 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 $\psi$ if the $z$-score of $z(x, \psi)$ is in the native range.

The starting point for our discussion is a potential function $W(x, \psi)$ evaluating the energy of a sequence $x$ when folded into a structure $\psi$ which is defined by the spatial coordinates of its $C^o$ and $C^\beta$ atoms, respectively. In this contribution we
consider two potential functions based on quite different encoding of the protein structures.

**Sippl’s PROSA II Potential** is a true pair potential with an additional surface term. It was designed to evaluate experimentally determined protein structures, to identify incorrectly folded proteins (or sections of proteins), and as an independent method for evaluating theoretical models of protein structures (Casari and Sippl, 1992; Hendlich et al., 1990; Sippl, 1990, 1993a,b). It is of the form

\[ W_s(x, \psi) = \sum_{i<j} W_s[x_i, x_j, |i - j|, d_{ij}] + \sum_i V_s[x_i; \chi(i)]. \]  

(1)

The additive pair-contributions \( W_s[a, b, k; r] \) depend on the type \( \gamma = C^\alpha \) or \( C^\beta \) 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 \( r = d_{ij} \) between the backbone atoms. The surface term \( V_s[a; \chi] \) depends on the type \( \gamma \) of the backbone atom, the amino acid \( a = x_i \) at sequence position \( i \) and the number \( \chi \) 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 (Bowie et al., 1990, 1991; Luthy et al., 1992). The parameter \( \chi \) 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 \( C^\beta \) backbone.

**The Neural Network (NN) Potential** includes multi-body interactions (Grossman et al., 1995). 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 \( C_\alpha \) 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 \( C_\alpha \) 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 \( C_\alpha \). Any residue within an interaction radius of e.g. 8Å to any \( C_\alpha \) atom can be labeled with invariant \( x, y, z \) coordinates. To solve the problem of how to usefully order the list of co-ordinates of the spatial neighbors the sphere surrounding each \( C_\alpha \) 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.
Table 1. Average length $\ell$ of Adaptive Walks to reach wildtype $z$-score. The length of the walk $\ell$ is averaged over 5 runs. $\ell/n$ is the average length walk normalized by the number of amino acids $n$ in each sequence.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PROSA II $\ell$</th>
<th>PROSA II $\ell/n$</th>
<th>NN $\ell$</th>
<th>NN $\ell/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1cbn</td>
<td>18.7</td>
<td>0.406</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1ubq</td>
<td>61.9</td>
<td>0.814</td>
<td>75.4</td>
<td>0.992</td>
</tr>
<tr>
<td>1adr</td>
<td>31.7</td>
<td>0.417</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>4icb</td>
<td>60.3</td>
<td>0.793</td>
<td>75.4</td>
<td>0.992</td>
</tr>
<tr>
<td>2trxa</td>
<td>71.7</td>
<td>0.664</td>
<td>112.0</td>
<td>1.037</td>
</tr>
<tr>
<td>1rrb</td>
<td>79.1</td>
<td>0.732</td>
<td>125.6</td>
<td>1.163</td>
</tr>
<tr>
<td>1ceu</td>
<td>44.1</td>
<td>0.408</td>
<td>76.0</td>
<td>0.703</td>
</tr>
<tr>
<td>1lyz</td>
<td>58.2</td>
<td>0.451</td>
<td>115.2</td>
<td>0.893</td>
</tr>
</tbody>
</table>

3. $z$-Scores, Adaptive Walks, and Inverse Folding

As a measure for the quality of fit of sequence $x$ and structure $\psi$ we use the $z$-score (Casari and Sippl, 1992)

$$z(x, \psi) = \frac{W(x, \psi) - \overline{W}(x)}{\sigma_W(x)}.$$  \hspace{1cm} (2)

Here $\overline{W}(x)$ is the average energy of sequence $x$ in all conformations in a database and $\sigma_W(x)$ is the standard deviation of the corresponding distribution. For the PROSA potential we use the same database as in (Casari and Sippl, 1992) [add half a sentence for NN potential]. 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 (Casari and Sippl, 1992). Gaps are not allowed in the alignments of sequences to structures with the potential functions used in this contribution.

Formally, we translated inverse folding into an optimization problem on the set of all sequences: we are looking for the optima $x$ of the $z$-score $z(x, \psi)$.

From the computational point of view, this optimization problem appears to be very easy. Indeed, it is sufficient to use the simplest heuristic, the adaptive walk, which repeatedly tries random mutations that are accepted if and only if the $z$-score improves, see Figure 1. We use only point-mutations in this study. 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 stop when a predefined threshold score $z^*$ is reached. In most cases we choose $z^*$ identical or slightly better than the $z$-score of the wildtype sequence/structure.

\footnote{The $z$-scores produced by PROSA II are the negative of the numbers quoted in this paper. We have inverted the sign to facilitate the comparison between the two potentials.}
Figure 1. Adaptive walks used for inverse folding of the 1cew structure with the NN potential (solid lines) and the PROSA II potential (dotted lines). The wildtype z-scores are indicated by the two dashed lines, the upper one for PROSA II, the lower one for the NN potential.

pair. In the case of the PROSA II potentials we require that both the C$^\alpha$ and the C$^\beta$ z-scores improve with each step of the adaptive walk.

Adaptive walks already yield some insight into the structure of protein space. The length $\ell$ of adaptive walks, that is, the number of accepted steps until $z^*$ is reached, gives information about the ruggedness of the energy landscape (Kauffman, 1993). Longer walks imply smoother surfaces with few local optima. In Table 1 it is shown that adaptive walks with the NN potential are consistently 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.


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 between the two potential functions, see Figure 2 and Table 2. However, 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.

It turns out that along an adaptive walk, we obtain an approximately linear relationship

\[
z' = az + b
\]

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. The data in Table 2 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 11yz with NN-optimized sequences, while the best PROSA-optimized sequences are still 0.8 z-score units short of the NN wildtype level.

The inaccuracies and inconsistencies of the two potentials are reflected by the fact that \( a_{NN/PROSA} \times a_{PROSA/NN} \gg 1 \) while, if the potentials were equivalent, we would observe that the product is exactly 1.
Table 2. z-score comparisons. The sequences encountered along adaptive walks performed with one potential were evaluated with the other potential, see equ. 3. \( z_{\text{opt}} \) denotes the best scores in the other potential, which should be compared to the corresponding wild type z-score \( z_{\text{wt}} \).

<table>
<thead>
<tr>
<th>Protein</th>
<th>( \rho )</th>
<th>slope</th>
<th>intercept</th>
<th>( z_{\text{opt}} )</th>
<th>( z_{\text{wt}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adaptive walks with NN potential</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4icb</td>
<td>0.819</td>
<td>0.24 \pm 0.005</td>
<td>3.20 \pm 0.06</td>
<td>8.1</td>
<td>10.06</td>
</tr>
<tr>
<td>1tro</td>
<td>0.932</td>
<td>0.31 \pm 0.003</td>
<td>2.17 \pm 0.04</td>
<td>8.8</td>
<td>10.88</td>
</tr>
<tr>
<td>1cew</td>
<td>0.938</td>
<td>0.25 \pm 0.002</td>
<td>1.86 \pm 0.03</td>
<td>7.9</td>
<td>6.20</td>
</tr>
<tr>
<td>2trxa</td>
<td>0.945</td>
<td>0.38 \pm 0.003</td>
<td>2.13 \pm 0.03</td>
<td>7.9</td>
<td>8.06</td>
</tr>
<tr>
<td>1lyz</td>
<td>0.919</td>
<td>0.37 \pm 0.003</td>
<td>1.25 \pm 0.04</td>
<td>7.5</td>
<td>7.70</td>
</tr>
<tr>
<td></td>
<td>Adaptive walks with PR( S ) II potential</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4icb</td>
<td>0.886</td>
<td>0.44 \pm 0.007</td>
<td>-1.40 \pm 0.07</td>
<td>6.6</td>
<td>8.08</td>
</tr>
<tr>
<td>1tro</td>
<td>0.938</td>
<td>0.49 \pm 0.004</td>
<td>-1.16 \pm 0.05</td>
<td>7.2</td>
<td>9.80</td>
</tr>
<tr>
<td>1cew</td>
<td>0.913</td>
<td>0.58 \pm 0.006</td>
<td>-2.57 \pm 0.06</td>
<td>6.9</td>
<td>5.91</td>
</tr>
<tr>
<td>2trxa</td>
<td>0.904</td>
<td>0.47 \pm 0.006</td>
<td>-1.21 \pm 0.06</td>
<td>7.3</td>
<td>8.06</td>
</tr>
<tr>
<td>1lyz</td>
<td>0.836</td>
<td>0.30 \pm 0.005</td>
<td>0.71 \pm 0.05</td>
<td>6.5</td>
<td>7.45</td>
</tr>
</tbody>
</table>

5. Neutral Sets

Let us now return to the structure of the neutral sets \( S(\psi) \) and their embedding in the sequence space \( \mathcal{Q}_n^\psi \) of all amino acid sequences of length \( n \). We define the neutral set of a native structure \( \psi \) to be the set of all sequences that fold into \( \psi \) according to our z-score criterion,

\[
S(\psi) = \{ x \in \mathcal{Q}_n^\psi | z(x, \psi) \geq z^* \}.
\]

We have seen already in a previous study (Babajide et al., 1997), and in section 3, that sequences generated by independent adaptive walks show little or no homology to the wild-type sequence or among each other. This is consistent with the observation that a significant sequence homology is not necessary for two proteins to have a common fold (Orengo et al., 1994).

The shape or topology of neutral sets has important implications for the evolution of proteins and for de novo design. For example, it has been frequently observed that seemingly unrelated sequences have essentially the same fold (Holm and Sander, 1997; Murzin, 1994, 1996). Whether these may have originated from a common ancestor, or whether they must be the result of convergent evolution, depends on the geometry of \( S(\psi) \) (Babajide et al., 1997).

The results of our adaptive walks already 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 3 show that the average Hamming distances \( \langle d \rangle_{\text{inv}} \) 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.
TABLE 3. Characteristics of Neutral Sets.

The average distances of inverse folded sequences (10 neutral walks and 5 adative walks for each structure), $\langle d \rangle_{adw}$, and the average distances between the endpoints of independent neutral walks, $\langle d \rangle_{nw}$, are comparable to the sequence length in both potentials.

<table>
<thead>
<tr>
<th>Protein</th>
<th>n</th>
<th>$\langle d \rangle_{adw}/n$</th>
<th>$\langle d \rangle_{nw}/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NN</td>
<td>PRoSA</td>
</tr>
<tr>
<td>1cbn</td>
<td>46</td>
<td>--</td>
<td>0.841</td>
</tr>
<tr>
<td>1ubq</td>
<td>76</td>
<td>0.896</td>
<td>0.803</td>
</tr>
<tr>
<td>1adr</td>
<td>76</td>
<td>--</td>
<td>0.718</td>
</tr>
<tr>
<td>4icb</td>
<td>76</td>
<td>0.539</td>
<td>0.731</td>
</tr>
<tr>
<td>2trxA</td>
<td>108</td>
<td>0.895</td>
<td>0.812</td>
</tr>
<tr>
<td>1rro</td>
<td>108</td>
<td>0.590</td>
<td>0.783</td>
</tr>
<tr>
<td>1ecz</td>
<td>108</td>
<td>0.510</td>
<td>0.762</td>
</tr>
<tr>
<td>1lyz</td>
<td>129</td>
<td>0.916</td>
<td>0.822</td>
</tr>
</tbody>
</table>

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 (Reidys et al., 1997), extensive sample statistics (Schuster et al., 1994) using neutral walks as a “probe”, and exhaustive folding of all sequences with given chain length $n$ (Grüner et al., 1996b).

The mathematical model assumes that sequences forming the same structure are distributed randomly using the fraction $\lambda$ of neutral neighbors as (the only) input parameter. If $\lambda$ is large enough this model makes two rather surprising predictions (Reidys et al., 1997; Reidys, 1997):

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

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

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 $\psi$, see the following section.

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 (Schuster et al., 1994; Babajide et al., 1997). The usefulness of this approach is demonstrated in (Göbel et al., 1997).
A neutral path starting at a sequence $x_0$ folding into a structure $\psi$ 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 $\psi$, and
(iii) the Hamming distance $d_H(x_0, x_i) = i$, i.e., each mutation increases the distance from the starting point (Schuster et al., 1994).

Since we have not solved the folding problem, we have to resort to a slightly weaker notion of neutrality, namely the one defined by the neutral sets $S(\psi)$, i.e., we accept a sequence $x_i$ as folding into the prescribed structure $\psi$ if its $z$-score is better than the wild-type score $z^*$. A neutral path ends after $L \leq n$ steps when no mutant of $x_L$ can be found that has Hamming distance $L + 1$ from the starting point and folds into $\psi$. In case of the PROSA II potential we require (as above) that this condition is satisfied independently for both the $C_\alpha$ and the $C_\beta$ potentials. In case of the NN potential we used equ.(7). Since we find that the PROSA $z$-scores of the accepted sequences along a neutral walk stay within the $z$-score interval $[z^*, z^*+1]$, see Figure 3, we have not used an upper bound on the $z$-scores in the PROSA case.

Inverse folded sequences with $z$-scores better than the threshold $z^*$ were used as starting points for neutral paths. Neutral walk experiments are summarized in Table 4.
TABLE 4. Neutral Walks. For each structure 10 neutral walks were performed. PROSA data for the structures marked with † taken from (Babajide et al., 1997).

<table>
<thead>
<tr>
<th>Protein</th>
<th>n</th>
<th>$\mathcal{L}$</th>
<th>$\mathcal{L}/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NN</td>
<td>PROSA</td>
</tr>
<tr>
<td>lcbn†</td>
<td>46</td>
<td>36.1</td>
<td>44.6</td>
</tr>
<tr>
<td>1ubq†</td>
<td>76</td>
<td>64.7</td>
<td>72.5</td>
</tr>
<tr>
<td>1adr</td>
<td>76</td>
<td>---</td>
<td>74.9</td>
</tr>
<tr>
<td>4icb</td>
<td>76</td>
<td>68.2</td>
<td>73.6</td>
</tr>
<tr>
<td>1rro</td>
<td>108</td>
<td>95.9</td>
<td>105.4</td>
</tr>
<tr>
<td>2trxA†</td>
<td>108</td>
<td>87.5</td>
<td>106.3</td>
</tr>
<tr>
<td>lcew</td>
<td>108</td>
<td>100.8</td>
<td>106.7</td>
</tr>
<tr>
<td>1lyz†</td>
<td>129</td>
<td>115.4</td>
<td>126.2</td>
</tr>
</tbody>
</table>

Figure 4. Schematic drawing of the protein structures used for closest approach experiments.

6. 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. More technically, we want to determine how close the neutral sets $S(\psi)$ and $S(\varphi)$ come together. This is of particular interest if $\psi$ and $\varphi$ of two unrelated shapes, one, say containing only $\beta$-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

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

exactly. The following computer experiment, however, can be used to provide good upper bounds on $D_f$. For the purpose of this section we have adopted a more restrictive definition of the neutral set of a shape $\psi$ by requiring that the z-score is contained in a narrow interval:

$$S(\psi) = \{x \in Q_20^n | 1.02z^* \geq z(x, \psi) \geq z^*\}$$
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.

For the computer experiment we need proteins of the same length. In the following we use three structures, all of which have length \( n = 108 \):

- **1ces** (Bode et al., 1988) is the 2Å resolution structure of a cystatin from chicken egg white. Cystatin is a phosphoprotein and a cysteine proteinase inhibitor belonging to the same superfamily as the stefin family and the kininogen family (Bode and Turk, 1991).

- Thioredoxin is an electron transport protein. It controls the activities of various enzymes in many kinds of cells by reducing disulfide bonds. The active form of thioredoxin contains two cysteine which are oxidized to form a disulfide bond when thioredoxin activates other enzymes. We use chain \( \alpha \) of the 1.68Å resolution structure 2trx (Katti et al., 1990) from E. Coli.

- The protein **irro** (Ahmed et al., 1993) is a rat oncomodulin isolated from morris hepatoma 5123TC. It is a small, acidic calcium-binding protein forming an all-\( \alpha \) structure. The **irro** structure has a resolution of 1.3Å.

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 4.

The residual distances (measured as Hamming distances) of the sequences resulting from closest approach walks, shown in Table 5, are surprisingly small even for pairs of protein structures that have virtually no structural features in common, see Figure 4.

### 7. JANUS as an Example

In a quite spectacular experiment Dalal et al. (Dalal et al., 1997) have designed a protein, JANUS that has 50% sequence identity with a predominantly \( \beta \)-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
Table 5. Closest approach walks

<table>
<thead>
<tr>
<th>Potential</th>
<th>( D_1 )</th>
<th>( D_1/n )</th>
<th>( D_1/n )</th>
<th>( D_1/n )</th>
<th>( D_1/n )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>icew/2trxa</td>
<td>4.9</td>
<td>0.045</td>
<td>0.665</td>
<td>0.699</td>
<td>0.648</td>
<td>20</td>
</tr>
<tr>
<td>NN</td>
<td>5.2</td>
<td>0.048</td>
<td>0.838</td>
<td>0.843</td>
<td>0.785</td>
<td>10</td>
</tr>
<tr>
<td>2trxa/1rro</td>
<td>6.5</td>
<td>0.060</td>
<td>0.755</td>
<td>0.711</td>
<td>0.711</td>
<td>13</td>
</tr>
<tr>
<td>NN</td>
<td>6.5</td>
<td>0.060</td>
<td>0.858</td>
<td>0.834</td>
<td>0.801</td>
<td>10</td>
</tr>
</tbody>
</table>

fold, usually by replacing them with the corresponding amino acid from Rop (only 9 positions in Janus differ from both Rop and 1pgb). In the Janus molecule the 7 amino acid tail found in Rop was retained to increase solubility.

Figure 5. Schematic drawing of the structures of 1rop and 1pgb. Also shown are the wildtype sequences, as well as the Janus sequence designed by Dalal et al. (Dalal et al., 1997).

We have 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 wildtype 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).

Both copies of the 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 wildtype z-score of 1rop. When using the PROSA potential we require that both Cα- and Cβ-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 6. 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. Jamus 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 sequence 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 Jamus sequence of 41%.

Several sequence positions were conserved in all or most of our simulations and also agreed with the amino acids chosen by Dalal et al. (Dalal et al., 1997). 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.

A comparison of z-scores for 1pgb, 1rop, the Dalal et al. designed sequence (jamus-dalal), and the average of nine of our evolved sequences (jamus-simulation), is shown in Table 6.
Table 6. $z$-score evaluation of wildtype and Janus sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Structure 1rop</th>
<th>Structure 1pgb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1pgb</td>
<td>-1.61</td>
<td>7.62</td>
</tr>
<tr>
<td>Janus (Dalal)</td>
<td>6.17</td>
<td>2.36</td>
</tr>
<tr>
<td>Janus (simulation)</td>
<td>7.52</td>
<td>2.00</td>
</tr>
<tr>
<td>1rop</td>
<td>7.47</td>
<td>0.72</td>
</tr>
</tbody>
</table>

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.

8. Discussion

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.

Neutral paths within the sets $S(\phi)$ extend to almost the length of the amino acid sequence. The neutral sets therefore form extensive neutral networks that percolate the entire sequence space. The existence of extensive neutral networks meets a claim raised by Maynard-Smith (Maynard-Smith, 1970) for protein spaces that are suitable for efficient evolution. Empirical evidence for a large degree of functional neutrality in protein space is indeed observed (Martínez et al., 1996).

Like the existence of neutral networks, shape space covering was first observed in computational studies of the RNA sequence structure relationship (Schuster et al., 1994; Grünert et al., 1996a,b); 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 (Sander and Schneider, 1991) 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.

The evolutionary implications of neutral networks and shape space covering discussed in detail in (Huynen et al., 1996; Huynen, 1996). 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. Empirical evidence for a large degree of functional neutrality in protein space was presented recently by Wain-Hobson and co-workers (Martínez et al., 1996).
Acknowledgements

This work was supported in part by the Austrian Jubiläums Fonds der Österreichischen National Bank, Proj. No. 6792, a travel grant from the New Mexican Diversity Biotechnology Consortium. A.B. was supported in part through a research grant from the Faculty of Sciences of the University of Vienna. A.S.L. and R.M.F. were supported by the Department of Energy under contract number W-7405-ENG-36.

References


*Proteins 17*: 355-362. URL: 
http://lore.came.sbg.ac.at/Extern/software/Prosa/prosa.html.