Exploring Protein Sequence Space Using Knowledge Based Potentials

ADERONKE BABAJIDE[a], ROBERT FARBER[c,d] IVO L. HOFACKER[a], JEFF INMAN[b,d], ALAN S. LAPEDES[c,d], AND PETER F. STADLER[a,d]

[a]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52742 Fax: **43 1 4277 52793

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

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

[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 numbers 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 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 (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; Hendlich *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 ψ 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 ψ which is defined by the spatial coordinates of its C^α and C^β 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

$$W_\gamma(x,\psi) = \sum_{i<j} W_\gamma[x_i, x_j, |i-j|; \mathbf{d}^\gamma_{ij}] + \sum_i V_\gamma[x_i; \chi(i)]\,. \tag{1}$$

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 $r = \mathbf{d}^\gamma_{ij}$ 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 (Bowie *et al.*, 1990, 1991; Luthy *et al.*, 1992). 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 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_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 co-ordinates 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.

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

	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

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

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

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

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[1]. 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

[1]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.

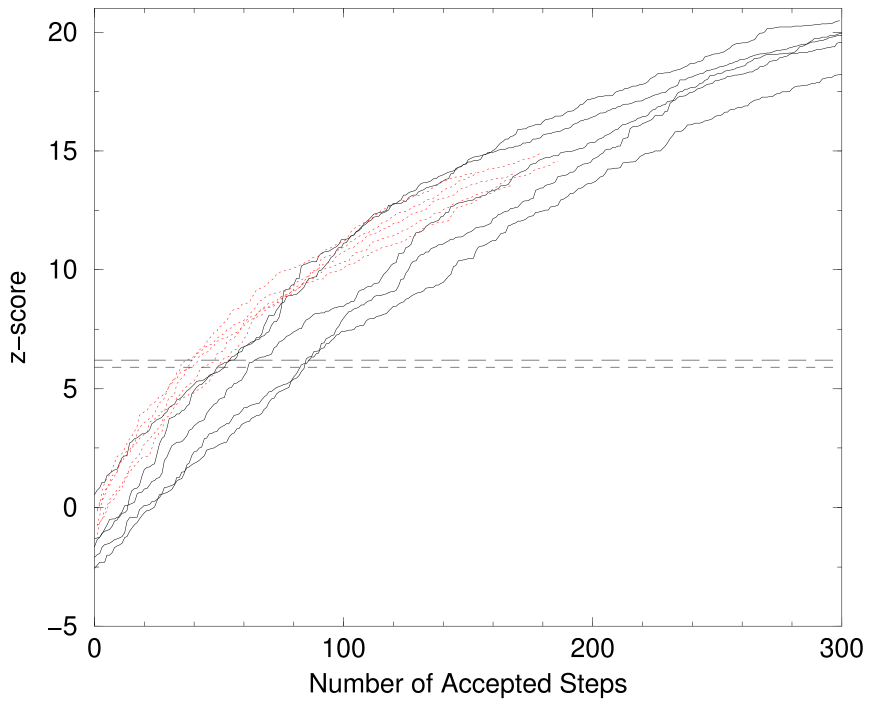


FIGURE 1. Adaptive walks used for inverse folding of the `1cew` structure with the `NN` potential (solid lines) and the `PROSA II` potential (dotted lines). The wildtype z-scores are indicated by the two dashed lines, the upper one for `PROSA II`, the lower one for the `NN` potential.

pair. In the case of the `PROSA II` potentials we require that both the C^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 (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.

4. Comparing the Two Potential Function Using Adaptive Walks

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.

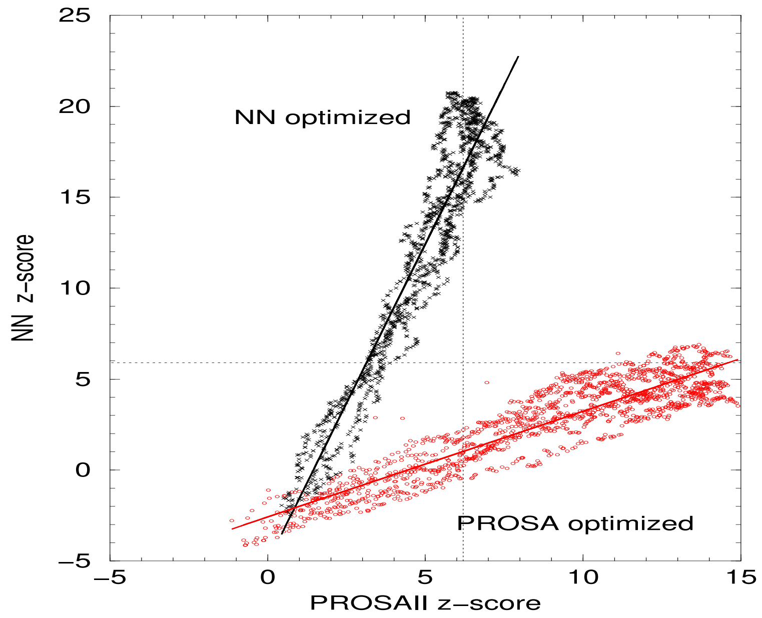


FIGURE 2. Correlation between `NN` and `PROSA II` z-scores for the `1cew` structure. The sequences encountered along the adaptive walks with one potential were re-evaluated with the other potential.

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

$$z' = az + b \tag{3}$$

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 `1lyz` 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 $a_{\mathtt{NN/PROSA}} \times a_{\mathtt{PROSA/NN}} \gg 1$ while, if the potentials were equivalent, we would observe that the product is exactly 1.

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_{opt} denotes the best scores in the other potential, which should be compared to the corresponding wild type z-score z_{wt}.

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
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	8.06
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
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

5. Neutral Sets

Let us now return to the structure of the neutral sets $S(\psi)$ and their embedding in the sequence space $\mathcal{Q}_{20}^n$ of all amino acid sequences of length n. We define the *neutral set* of a native structure ψ to be the set of all sequences that fold into ψ according to our z-score criterion,

$$S(\psi) = \{x \in \mathcal{Q}_{20}^n | z(x,\psi) \geq z^*\}\,. \tag{4}$$

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 adpative walksalready 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 $\langle d \rangle_{\text{adw}}$ 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.

TABLE 3. Characteristics of Neutral Sets.
The average distances of inverse folded sequences (10 neutral walks and 5 adative walks for each structure), $\langle d\rangle_{\rm adw}$, and the average distances between the endpoints of independent neutral walks, $\langle d\rangle_{\rm nw}$, are comparable to the sequence length in both potentials.

Protein	n	$\langle d\rangle_{\rm adw}/n$		$\langle d\rangle_{\rm nw}/n$	
		`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

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.*, 1996a).

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 (Reidys *et al.*, 1997; Reidys, 1997):

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

$$\lambda_{cr}(\alpha) = 1 - \sqrt[\alpha-1]{\frac{1}{\alpha}} \approx 0.146 \tag{5}$$

  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.

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).

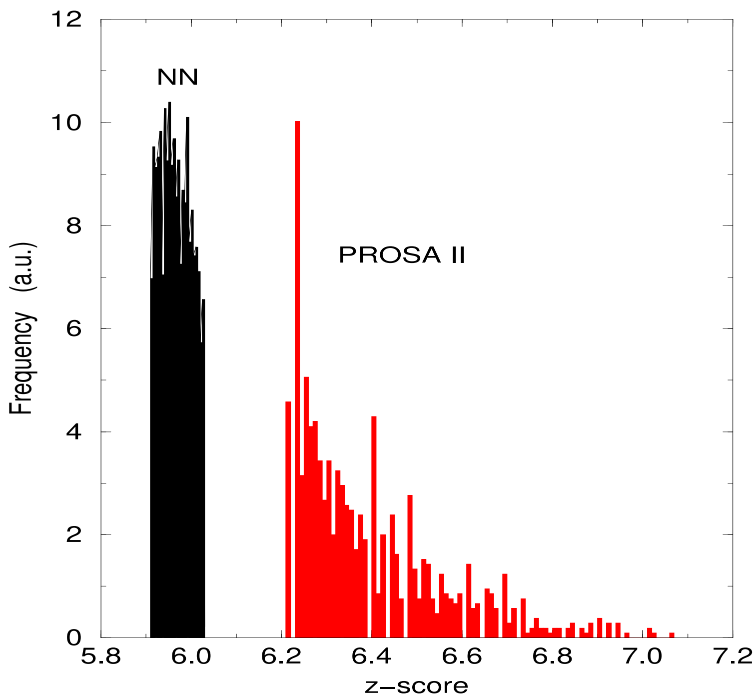


FIGURE 3. Distribution of z-scores encountered along neutral walks for the `1cew` shape. While there is no explicit upper bound on the acceptable `PROSA` z-scores, we find that the neutral walks are confined between wild-type level and only 1 z-score unit better than wildtype level.

A neutral path starting at a sequence x_0 folding into a structure ψ consists of sequences $x_1, x_2, \dots$ 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(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 ψ if its z-score is 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. In case of the `PROSA II` potential we require (as above) that this condition is satisfied independently for both the C_α and the C_β 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).

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

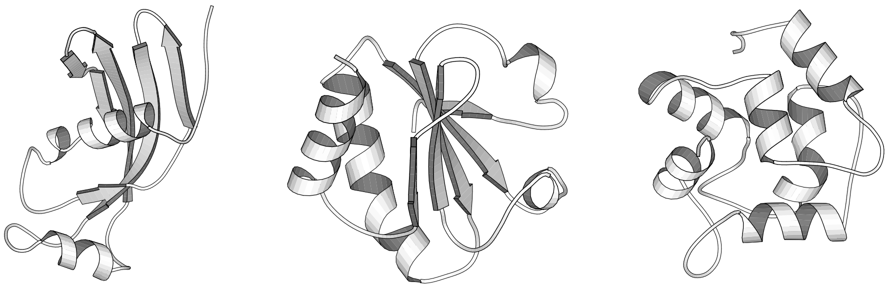


FIGURE 4. Schematic drawing of the protein structures used for closest approach experiments

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 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*

$$D_f(S(\psi), S(\varphi)) = \min\{d(x, x')|x \in S(\psi) \text{ and } x' \in S(\varphi)\} \tag{6}$$

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 ψ by requiring that the z-score is contained in a narrow interval:

$$S(\psi) = \{x \in \mathcal{Q}_{20}^n | 1.02z^* \geq z(x, \psi) \geq z^*\} \tag{7}$$

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 $n = 108$:
`1cew` (Bode *et al.*, 1988) is the 2 Å 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 `A` of the 1.68 Å resolution structure `2trx` (Katti *et al.*, 1990) from E. Coli.
The protein `1rro` (Ahmed *et al.*, 1993) is a rat oncomodulin isolated from morris hepatoma 5123TC. It is a small, acidic calcium-binding protein forming an all-α structure. The `1rro` structure has a resolution of 1.3 Å.

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.

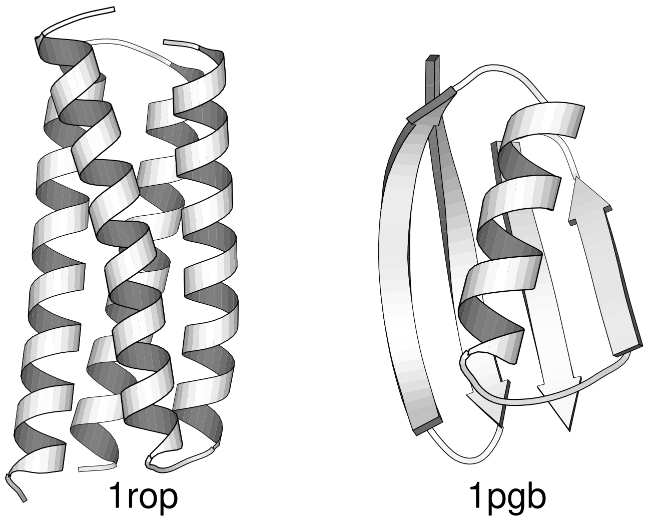
7. `JANUS` as an Example

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 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

TABLE 5. Closest approach walks

Potential	D_f	D_f/n	D_1/n	D_2/n	$\overline{D_f}/n$	N
1cew/2trxa						
PROSAII	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						
PROSAII	6.5	0.060	0.755	0.711	0.711	13
NN	6.5	0.060	0.858	0.834	0.801	10

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.



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

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

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

FIGURE 6. Nine of the 18 PROSA-evolved `janus` sequences. In the consensus line below '*', ':', and '.' mark conserved or nearly conserved positions.

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 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 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: `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 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 Janus 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 (janus-dalal), and the average of nine of our evolved sequences (janus-simulation), is shown in Table 6.

TABLE 6. z-score evaluation of wildtype and `janus` sequences

	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

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 $S(\psi)$ 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 (Martinez *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; Grüner *et al.*, 1996b,a): 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 (Martinez *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

Ahmed FR, Rose DR, Evans SV, Pippy ME, To R. 1993. Refinement of recombinant oncomodulin at 1.30Å resolution. *J Mol Biol 216*: 1216–1224.

Babajide A, Hofacker IL, Sippl MJ, Stadler PF. 1997. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding Design 2*: 261–269. Santa Fe Institute Preprint 96-12-085.

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

Bode W, Engh R, Musil D, Thiele U, Huber R, Karshikov A, Brzin J, Kos J, Turk V. 1988. The 2.0Å x-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J 7*: 2593–2599.

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

Bowie JU, Clarke ND, Pabo CO. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins 7*: 257–264.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*: 164–170.

Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potentials — hydrophobic potentials derived from X-ray structures of globular proteins is able to identify native folds. *J Mol Biol 224*: 725–732.

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

Dalal S, Balasubramanian S, Regan L. 1997. Protein alchemy: Changing β-sheet into α-helix. *Nat Struct Biol 4*: 548–552.

Drexler KE. 1981. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc Natl Acad Sci USA 78*: 5275–5278.

Göbel U, Forst CV, Schuster P. 1997. Structural constraints and neutrality in RNA. In: *Proceedings of the German Conference on Bioinformatics 1996*, Hofestädt R, Lengauer T, M. Löffler DS, eds. Berlin, Heidelberg, New York: Springer Verlag, volume 1278 of *Lecture Notes in Computer Science*, pp. 156–165.

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

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

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

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996a. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath Chem 127*: 375–389.

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996b. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monatsh Chem 127*: 355–374.

Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. 1990. 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.

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

Huynen MA. 1996. Exploring phenotype space through neutral evolution. *J Mol Evol 43*: 165–169.

Huynen MA, Stadler PF, Fontana W. 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci USA 93*: 397–401.

Katti SK, LeMaster DM, Eklund H. 1990. Crystal structure of thioredoxin from escherichia coli at 1.68åresolution. *J Mol Biol 212*: 167–184.

Kauffman SA. 1993. *The Origin of Order*. New York, Oxford: Oxford University Press.

Luthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*: 83–85.

Martinez MA, Pezo V, Marlière P, Wain-Hobson S. 1996. Exploring the functional robustness of an enzyme by *in vitro* evolution. *EMBO J 15*: 1203–1210.

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

Murzin AG. 1994. New protein folds. *Curr Opin Struct Biol 4*: 441–449.

Murzin AG. 1996. Structural classification of proteins: new superfamilies. *Curr Opin Struct Biol 6*: 386–394.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature 372*: 631–634.

Reidys C, Stadler PF, Schuster P. 1997. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull Math Biol 59*: 339–397.

Reidys CM. 1997. Random induced subgraphs of generalized n-cubes. *Adv Appl Math 19*: 360–377.

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

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

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: A case study in RNA secondary structures. *Proc Roy Soc London B 255*: 279–284.

Sippl MJ. 1990. 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.

Sippl MJ. 1993a. 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.

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