Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force
Aderonke Babajide¹, Ivo L Hofacker², Manfred J Sippl³ and Peter F Stadler¹,⁴

Background: Many protein sequences, often unrelated, adopt similar folds. Sequences folding into the same shape thus form subsets of sequence space. The shape and the connectivity of these sets have implications for protein evolution and de novo design.

Results: We investigate the topology of these sets for some proteins with known three-dimensional structure using inverse folding techniques. First, we find that sequences adopting a given fold do not cluster in sequence space and that there is no detectable sequence homology among them. Nevertheless, these sequences are connected in the sense that there exists a path such that every sequence can be reached from every other sequence while the fold remains unchanged. We find similar results for restricted amino acid alphabets in some cases (e.g. ADLG). In other cases, it seems impossible to find sequences with native-like behavior (e.g. QLR). These findings seem to be independent of the particular structure considered.

Conclusions: Amino acid sequences folding into a common shape are distributed homogeneously in sequence space. Hence, the connectivity of the set of these sequences implies the existence of very long neutral paths on all examined protein structures. Regarding protein design, these results imply that sequences with more or less arbitrary chemical properties can be attached to a given structural framework. But we also observe that designability varies significantly among native structures. These features of protein sequence space are similar to what has been found for nucleic acids.

Introduction
The number of possible proteins is enormous. For n = 100 residues there are 2ⁿ¹⁰⁰ sequences. On the other hand, the repertoire of stable native folds seems to be highly restricted or even vanishingly small [1,2]. Some obvious questions are: How many sequences have stable structures? How many sequences adopt the same fold? How are these sequences distributed in sequence space?

Sequences belonging to the same fold ψ form a subset $S(\psi)$ of sequence space. These sequences will be 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. For example, it has been frequently observed that seemingly unrelated sequences have essentially the same fold [2–4]. Whether these may have originated from a common ancestor, or whether they must be the result of convergent evolution, depends on the geometry of $S(\psi)$ (see Figure 1).

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 $S(\psi)$, i.e. 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 [5,6].

The native structure of a given amino acid sequence corresponds to the minimum of its free energy. If this energy function 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 can indeed be used to investigate the sequence/structure relation for RNA molecules [7].

Inverse folding is, however, 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 (groundstate) of each sequence is equal to 0 (or any other constant). 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.
Recent studies using knowledge-based potentials [6,8–14] have demonstrated that the energy of the native fold (i.e., putative groundstate) 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 is a member of $\mathcal{S}(\psi)$ if the z-score of x in conformation is in the native range.

In this paper, we describe our computational procedure in detail and show that the PROSA II potentials [13,14] are a suitable tool for studying neutral sets in the sequence space of proteins. Then we report that neutral sets contain sequences with no noticeable homology and that there are neutral paths that connect seemingly unrelated sequences.

### Results

#### The distribution of inverse folded sequences

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 [15]. The distribution of pairwise Hamming distances for 700 sequences with z-score approximately −11 on the 2TRX structure is shown in Figure 2a. Although they lie somewhat closer together than random sequences with a typical amino acid composition (taken from the Swiss-Prot database), pairs with the maximum Hamming distance, $n = 108$, do occur. Note that maximum Hamming distance implies that all positions of two sequences are occupied by distinct amino acids. Also, we find that the distribution of pair distances depends only slightly on the threshold z-score, $z^*$ (data not shown).

The distribution of Hamming distances of sequences taken from different adaptive walks has been analyzed in detail for traces of clustering. Tree reconstruction methods, such as neighbor joining (see Figure 2b), and the split decomposition technique [16] suggest that the sequences with wild-type-like z-scores are randomly distributed in sequence space. The average Hamming distance $\langle d \rangle_{ave}$ of the endpoints of independent walks is ~80–90% of the chain length, significantly below the expected value of 95% in a 20-letter alphabet. This indicates some common features among these sequences that are not detectable by sequence comparison. The $\langle d \rangle$ data compiled in Table 1 can be explained, however, by the non-uniform distribution of amino acid frequencies.

A number of groups have argued that the pattern of hydrophobic versus hydrophilic amino acids (HP-pattern) has a dominating influence on protein structure [17–19]. An example of the HP-pattern of inverse folded 2TRX sequences is shown in Figure 3.

There are only a very small number of positions in which only hydrophilic or only hydrophobic residues occur in all the sequences sampled from both inverse folding and neutral path (see below). The pattern shown in Figure 3 is typical: while for most positions the type of amino acid generally conforms to the wild type, the deviations are substantial. The thioredoxin wild-type sequence contains 72 hydrophobic residues (66.7%); the fraction of hydrophobic residues in Figure 3 is 69.7%. We find an average Hamming distance of 33.0 between the patterns in Figure 3, while the average distance between random sequences with the same fraction of hydrophilic residues is ~45.6. Furthermore, there are 14 conserved positions. Taking these into account, we would expect a Hamming distance of ~40 if the
Figure 2

(a) Distribution of pairwise Hamming distances for 700 sequences designed to have z-scores of about -11 on the 2TRX structure (unbroken line) and for 500 random sequences with typical amino acid composition (broken line). The vertical line at 95.15 is the average distance to the 2TRX sequence. Its position close to the maximum of the distribution indicates that the wild-type sequence is not exceptional among the inverse folded sequences. (b) A neighbor-joining analysis of 23 sequences from unrelated adaptive walks with z-scores slightly smaller than -11. The data do not indicate a significant amount of clustering.

remaining positions were uncorrelated. We conclude that inverse folding is very flexible at the level of individual amino acids but requires a significant level of conservation of amino acid classes.

Table 1

<table>
<thead>
<tr>
<th>Characteristics of neutral networks.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB ID</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1CBN</td>
</tr>
<tr>
<td>1UBQ</td>
</tr>
<tr>
<td>2TRXA</td>
</tr>
<tr>
<td>1LYZ</td>
</tr>
</tbody>
</table>

*The length of neutral path is averaged over all data with z-scores between wild type and three standard deviations better than wild type.

The existence of neutral networks in protein space

A ‘neutral path’ starting at a sequence $x_0$ folding into a structure $\psi$ consists of sequences $x_1, x_2, \ldots$ such that the sequence $x_i$ is obtained by a single point mutation from $x_{i-1}$ for all $i > 0$, all sequences $x_i$ fold into $\psi$, and the Hamming distance $d_H(x_0, x_i) = i$, i.e. each mutation increases the distance from the starting point [20]. 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 if its z-score, as well as the z-scores computed separately from the $C_{\alpha}$ and the $C_{\beta}$ potentials, is below a threshold value $z^*$. This value is chosen at least as low as the wild-type z-score. A neutral path therefore ends after $z \leq n$ steps when no mutant of $x_0$ can be found that has Hamming distance $z + 1$ from the starting point and folds into $\psi$.

Figure 3

HP-patterns of sequences generated by independent neutral paths. +, Hydrophilic amino acids (R, N, D, E, Q, G, H, K, P, S, T and Y); −, hydrophobic amino acids (A, C, I, L, M, F, W and V). The first line of stars indicates the positions in which only hydrophobic or only hydrophilic residues occur. It is followed by the consensus sequence (cs). The last line (wt) shows the HP-pattern of the wild-type sequence. The stars between wild-type and consensus sequence highlight the overlap of these two sequences.
Inverse folded sequences with z-scores below the threshold $z^*$ were used as starting points for neutral paths. The substitution frequencies for the production of mutants were computed from the natural frequencies of the amino acids as contained in the Swiss-Prot database. Figure 4 shows the results for four different protein structures. We find that the lengths of the neutral paths $\mathcal{L}$ are roughly equal to the lengths of the proteins, at z-score levels comparable to the wild-type sequence.

Even at $z$-scores about six standard deviations better than the wild-type $z$-score, the length of the neutral path is still greater than three-quarters of the length of the protein. The average values of $\mathcal{L}$ taken over the $z$-score interval $z_{\text{thr}} - 3 \leq z \leq z_{\text{thr}}$ are collected in Table 1. The average Hamming distances between the endpoints of an unrelated neutral path, $\langle d \rangle_{\text{ent}}$, are in the range 90–95% of the chain length, indicating that the neutral networks span essentially the entire sequence space. It is not surprising that the Hamming distances between the endpoints of neutral paths are somewhat larger than the average distance between the endpoints of adaptive walks, since a neutral path has a built-in bias towards sequences that contain a more uniform distribution of amino acids.

**Restricted alphabets**

It is natural to ask whether all 20 amino acids are in fact necessary to build native protein structures, or whether this can already been done with a (small) subset of different amino acids. We have therefore attempted to find inverse folded sequences for our four test protein structures that contain only a small subset of amino acids. Not surprisingly, no sequences with wild-type-like structures could be found when only hydrophilic amino acids or only
hydrophobic amino acids were used. Surprisingly, however, we observed substantial differences between different alphabets that all contain both hydrophilic and hydrophobic amino acids.

For instance, the two-letter alphabet AD gives very poor results, while other combinations of just one hydrophilic and one hydrophobic amino acid, such as LS or DL, yield wild-type-like z-scores for all four of our test proteins. These sequences in general also yield native-like secondary structures with SOPM or PHD. It is hard to believe that native protein folds such as 1LYZ or 2TRXA can be formed from a two-letter amino acid alphabet. However, Kamtekar et al. [18] observed that “a simple binary code of polar and non-polar residues arranged in the appropriate order can drive polypeptide chains to collapse into globular α-helical folds.”

It is not surprising that ADL and ADLG yield good sequences, since DL is already sufficient. The inverse folded sequences in these alphabets do, however, contain a substantial fraction of A and G. The alphabet ADLG has been proposed as a candidate for a primordial set of amino acids, before the full genetic code was developed [21]. It is reassuring to see that this alphabet allows inverse folding of a variety of present day protein structures. It is worth noting in this context that the QLR alphabet used in experimental work on random polypeptides by Sauer and co-workers [22,23] does not yield wild-type-like z-scores for globular proteins. This may not be surprising, since Sauer’s experimental QLR peptides form multimeric structures.

Extensive studies on neutral paths have been performed only for the ‘primordial alphabet’ ADLG. The average length of neutral path $\mathcal{L}$ and the fraction $\mathcal{L}/n$ for our test proteins are: crambin (1CBN), $\mathcal{L} = 40.2$, $\mathcal{L}/n = 0.87$; ubiquitin (1UBQ), $\mathcal{L} = 58.0$, $\mathcal{L}/n = 0.76$; thioredoxin (2TRX), $\mathcal{L} = 87.5$, $\mathcal{L}/n = 0.81$; and lysozyme (1LYZ), $\mathcal{L} = 109.0$, $\mathcal{L}/n = 0.84$. Note that the distance between random sequences is $0.75n$ in a four-letter alphabet. The length of neutral paths thus extends well beyond the mean distance of random sequences even in the highly restricted ADLG alphabet.

The best z-scores relative to the wild type have the order 1CBN > 1LYZ > 2TRXA > 1UBQ irrespective of the amino acid alphabet (see Figure 5). We find the same ordering for $\langle d \rangle_{\text{wild}}/n$ and $\langle d \rangle_{\text{adj}}/n$ while the relative lengths of $L_{\text{adj}}/n$ give the reverse order. This suggests that ‘designability’ is indeed an intrinsic property of the structure and that it may vary significantly among native structures [24–26].

**Discussion**

**Investigation of the neutral subsets $S(\psi)$ of protein space using inverse folding yields global properties of the sequence/structure relations of proteins**

We found that sequences with native-like z-scores can indeed be constructed using simple adaptive walks in sequence space. Also, more sophisticated optimization techniques are not necessary, since local minima of z-score in the high dimensional sequence space are rare. We found essentially no homology between the inverse folded sequences and no discernible clustering. The distribution

![Figure 5](image-url)

Inverse folding with restricted alphabets. The data clearly show that some restricted alphabets allow for inverse folding of native structures while we could not find acceptable z-scores with other alphabets. It is also interesting to note that the quality of best solutions depends on the structure under consideration, and that this dependence is the same for all restricted alphabets that we have investigated. One might speculate that there are in fact protein structures that are attained by a much larger number of sequences than other structures.
of amino acids in these sequences is essentially random. Of course, the sequences belonging to a neutral set \( S(\psi) \) do have important features in common, such as a native-like z-score and very similar predicted secondary structures.

The neutral paths within the sets \( S(\psi) \) extend to the length of the amino acid sequence at z-scores comparable to the wild-type score. The vast extension of the network of neutral paths suggests that extensive neutral networks of sequences folding into the same structure percolate the entire sequence space [27]. The same qualitative results were obtained for some, but not all, restricted amino acid alphabets that contain both hydrophilic and hydrophobic amino acids. The existence of extensive neutral networks meets a claim raised by Maynard-Smith [28] for protein spaces that are suitable for efficient evolution. The evolutionary implications of neutral networks are explored in detail in [29,30]. Empirical evidence for a large degree of functional neutrality in protein space was presented recently by Wain-Hobson and co-workers [31].

In this work, we have restricted ourselves to sequences of constant length to avoid the problem of defining when two structures of different length are considered equal. Allowing insertion and deletion in addition to point mutations should, if anything, increase the extent of neutral paths, since it increases the connectivity of sequence space.

It is interesting to compare these findings with the properties of the sequence/structure map of RNA molecules. Secondary structures of nucleic acids, i.e. the patterns of Watson–Crick and GU base pairs, are used on a regular basis to describe and interpret experimental results, they are conserved in evolution, and, last but not least, there are efficient algorithms for computing them [32–35]. In a series of extensive computational studies, the most salient features of the ‘combinatory map’ of RNA secondary structures have been elucidated [20,36–43] and a number of sometimes unexpected global properties have been discovered:

1. There are many more sequences than secondary structures, hence many sequences fold into the same structure.
2. The distribution of the number of sequences folding into the same structure is highly non-uniform, i.e. there are few very common structures and many different very rare structures.
3. The sequences folding into a common structure are distributed randomly in sequence space. No clustering is visible.
4. The sequences folding into a common structure form extended neutral networks.
5. The average distance from a random sequence to a sequence that folds into a desired structure is short compared to the maximum distance in sequence space.

Our results draw a similar picture for protein space. One issue that could not be addressed in the present study concerns shape space covering. RNA does exhibit shape space covering, i.e. any common structure can be found within a small radius in sequence space centered around an arbitrary reference point [20,41,42]. It is not clear at this point whether protein (or the knowledge-based potential model of protein space considered here) shares this property. Sander and Schneider [44] have argued that sequences with more than 30% sequence homology will give rise to the same fold. Random graph models of neutral networks [27], on the other hand, predict that the neutral networks of any two different structures should come close at least in some parts of sequence space. Further computer experiments will be necessary to decide whether protein space exhibits shape space covering or whether there is a subtle form of clustering that we could not detect in the limited set of inverse folded sequences.

**Materials and methods**

**Potential function**

'Sequence space' is the set of all possible sequences of a given length together with a distance measure, usually the Hamming distance [45], i.e. the number of positions with different amino acids. For the space of amino acid sequences, we shall use the synonym 'protein space' [28].

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 that is defined by the spatial coordinates of its \( C_{\alpha} \) and \( C_{\beta} \) atoms, respectively. We use the software package PROSA II, which has been 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 [12–14,46,47]. The potential functions used in this program take the form:

\[
W(x,\psi) = \sum_{\gamma < j} W_{\gamma}[(x_i, x_j) - \beta d_{ij}] + \sum_{\gamma j} V_{\gamma}(x_i, x_j) \tag{1}
\]

The additive pair contributions \( W_{\gamma}(a,b,k,l) \) 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 \( \varepsilon = d_{ij} \) of the backbone atoms. The surface term \( V_{\gamma}(a,b) \) 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 [6,48,49]. The parameter \( \chi \) serves as a (crude) quantitative measure for the surface exposure of residue \( a \).

As a measure for the quality of fit of sequence \( x \) and structure \( \psi \) we use the z-score [46]:

\[
z(x,\psi) = \frac{W(x,\psi) - \bar{W}(x)}{\sigma_W(x)} \tag{2}
\]

Here, \( \bar{W}(x) \) 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. We use the same database as in [46]. Normalization of energies is necessary, since the relative groundstate energies of different sequences are not available. The z-score introduces a proper normalization, where the range of values of native folds is known [46].
The z-scores of perturbed protein structures. A set of perturbed protein structures were produced by 'heating' the structure in a molecular dynamics simulation and then cooling (and compactifying) it again (M Jaritz, personal communication). The data have been obtained using the $C^\alpha$ only potential of PROSA 1.0.

Figure 6 shows that the z-score correlates well with the rms deviation of alternative structures from the native structure (M Jaritz, personal communication). A second line of evidence comes from X-ray structures measured at different resolution by different groups (see Table 2). Their z-scores indeed improve with increasing resolution of the structure determination.

Inverse folding

We accept $x$ as a sequence folding into a structure $y$ if its z-score is at least as good as the z-score of the wild-type sequence $\hat{z}$. The z-score can thus be used as an approach to inverse folding; given a fixed backbone conformation, we may search for sequences $x$ that give z-scores $z(x,y)$ at least as low as the z-score of the native sequence $\hat{z}$. Of course, only native structures that are already in the database can be explored by this method. Formally, we have translated inverse folding into an optimization problem on the set of all sequences: we are looking for the minima $x$ of the z-score $z(x,y)$.

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 (exchanges of single amino acid) that are accepted if and only if the z-score decreases. While the procedure would eventually terminate in a local optimum, in practice we stop when a predefined threshold score $z^*$ is reached. The procedure is equivalent to a Monte-Carlo optimization, as used in [50,51], but at zero temperature. Typical examples of adaptive walks are shown in Figure 7.

The final sequences derived from adaptive walks on, for example, the 2TRX structure were up to 50% better than the score of the wild-type sequence ($z = 2.22$). However, there is no obvious evolutionary benefit in optimizing structural stability beyond a level that ensures correct folding, and hyperstability might even impede the folding process. We should therefore expect that, although sequences with z-scores far better than that of the wild type can be found, they do not correspond to native sequences. Therefore, we restrict the search for sequences to z-scores not more than 3 units below wild-type level.

The length $L_{\text{adap}}$ of an adaptive walk is defined as the number of accepted mutations that are necessary to reach the wild-type z-score. The ratio $L_{\text{adap}}/n$, where $n$ denotes the sequence length, indicates the relative abundance of sequences that fold into the wild-type structure, because rare structures can be reached only by very long walks (or not at all) while short walks should suffice to find a sequence folding into a very common structure. Numerical data can be found in Table 1.

Reliability

Whether the sequences predicted by our inverse folding procedure do indeed fold to the desired structure can ultimately be answered only by experiment. Independent criteria, however, would at least indicate whether the assumption is reasonable. Such a criterion is the predicted secondary structure compared to the known secondary structure of the target conformation. The best available algorithms combine secondary structure prediction with a search for homologous sequences and thereby attain accuracies over 70% for the assignment of residues to helix, strand and loop regions [52]. Since our inverse folded sequences have no or little homology to known sequences, we have to expect somewhat lower accuracies.

Table 2

<table>
<thead>
<tr>
<th>Resolution</th>
<th>$z$</th>
<th>$z_{\text{pair}}$</th>
<th>$z_{\text{surf}}$</th>
<th>PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0 Å</td>
<td>-13.56</td>
<td>-11.00</td>
<td>-9.37</td>
<td>1SPA</td>
</tr>
<tr>
<td>2.4 Å</td>
<td>-12.93</td>
<td>-10.18</td>
<td>-8.18</td>
<td>1AAW</td>
</tr>
<tr>
<td>2.5 Å</td>
<td>-12.77</td>
<td>-9.91</td>
<td>-8.16</td>
<td>3AAT</td>
</tr>
<tr>
<td>2.8 Å</td>
<td>-11.39</td>
<td>-7.54</td>
<td>-9.14</td>
<td>2AAT</td>
</tr>
</tbody>
</table>

It is interesting to note that the resolution affects, almost exclusively, the contribution of the pair potential, while it has only a negligible effect on the z-score contribution of the surface potentials.
We mostly used the program SOPM by Geourjon and Deleage [53]. It correctly predicts ~65% of the residues in the 2TRX wild type. Similar results were achieved using the PHD method of Rost and Sander [52]. Figure 8 shows the overlap between the 2TRX secondary structure and the SOPM prediction for every fifth sequence from the five adaptive walks shown in Figure 7. The overlap between the predicted and 2TRX secondary structure at first increases with improving z-score, then saturates at about 65% once the z-score becomes better than that of the wild-type sequence, as expected. Comparable results were obtained for the other structures. The PROSA potentials depend only on distances between Cα or Cβ atoms and surface exposure and make no explicit reference to secondary structure. The fact that the SOPM predictions of inverse folded sequences agree in general with the wild-type secondary structure indicates that these sequences indeed favor the 2TRX fold.

Acknowledgements
This research was supported in part by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project numbers 10578-MAT, 11205-MOB, and 11801-GEN, and a grant from the Diversity Biotechnology Consortium, New Mexico.

References


