Neutral Networks in Protein Space

A Computational Study Based on Knowledge-Based Potentials of Mean Force

ADERONKE BABAJIDE[a], IVO L. HOFACKER[b],
MANFRED J. SIPPL[c], AND PETER F. STADLER[a,d,*]

[a]Institut f. Theoretische Chemie, Univ. Wien, Austria
[b]The Beckman Institute, Univ. of Illinois, Urbana/Ill., U.S.A.
[c]CAME, Univ. Salzburg, Austria
[d]The Santa Fe Institute, Santa Fe, New Mexico, U.S.A.

[*]Mailing Address: Peter F. Stadler
Inst. f. Theoretische Chemie, Universität Wien
Währingerstr. 17, A-1090 Wien, Austria
Phone: **43 1 40480 665 Fax: **43 1 40480 660
E-Mail: `studla@tbi.univie.ac.at` or `stadler@santafe.edu`

Background: Protein space is explored by means of an inverse folding procedure that makes use of knowledge-based potentials of mean force.

Results: Computer simulations indicate that amino acid sequences folding into a common shape are distributed homogeneously forming extended percolating networks that span the entire sequence space.

Conclusions: The existence of very long neutral paths on all examined protein structures, indicates the existence of neutral networks percolating protein space. The same qualitative results were obtained for some, but not all, restricted amino acid alphabets. In this respect, the sequence-structure map of proteins seems to be very similar to the nucleic acid case.

Key words: inverse folding, knowledge based potentials, neutral mutations, restricted alphabets, sequence-structure map

1. Introduction

The prediction of protein conformations from amino acid sequences has remained an unsolved problem in spite of all efforts for a variety of reasons. Notably, existing potential functions are far too inaccurate to predict the correct ground state conformation and the computational difficulties of actually finding the groundstate conformation using a given potential function exceed the available computational resources by orders of magnitude. Thus accurate tertiary structure prediction is not yet possible; for a review of the state of the art see, e.g., [12]. As a consequence protein space beyond the few polypeptide structures that have been determined experimentally is largely *terra incognita*. Even basic global features of the sequence-structure relations of proteins remain in the dark.

The most interesting notion of *structure* in molecular biology, in particular in the context of evolution, does not refer to atomic coordinates but to an intermediate level of description that is rendered as ribbon or wire diagram. Lattice models and descriptions in terms of a contact matrix, such as RNA secondary structures, can be regarded as a special case of coarse grained representations. Extensive studies of lattice heteropolymer models by Ken Dill and co-workers [9, 10, 14, 53] and other groups [6, 7, 43] have focussed mostly on the uniqueness of ground states and questions of kinetic folding. A major drawback of these models is their computational difficulty [52], which precludes the study of large ensembles of sequences longer than some 20 monomers.

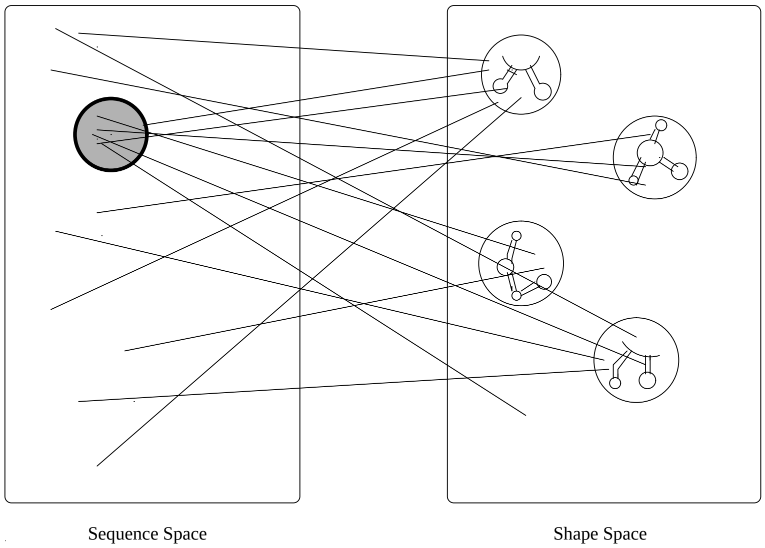

Figure 1: Sequence-structure relation in the case of RNA secondary structures: Sequences folding into a particular structure can be found anywhere in sequence space. Such sequences can be connected by extended networks of structurally neutral neighbors.

Secondary structures of nucleic acids 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 [28, 36, 54, 55]. In a series of extensive computational studies the most salient feature of the "combinatory map" of RNA secondary structures was elucidated over the past years [16, 17, 18, 42, 50, 29, 23, 24, 51, 30]. A number of sometimes unexpected global properties have been discovered:

- There are many more sequences than structures, hence many sequences fold into the same structure.

- The distribution of the number of sequences folding into the same structure is highly nonuniform, i.e., there are few very common structures and many different very rare structures.
- The sequences folding into a common structure are distributed randomly in sequence space. No clustering is visible.
- The sequences folding into a common structure form extended *neutral networks*, i.e., there are networks consisting of sequences that fold into the same structure which extend through all sequence space.
- 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.

Naturally, the question arises whether protein space has similar properties. Since protein folding is intractable at this time, we are forced to persue a less ambitious approach. Suppose we could solve the *inverse protein folding problem* [15, 5] that consists in finding a sequence that folds into a *given* protein conformation, or even less ambitiously, suppose we could reliably *decide whether a given sequence folds into a given structure*. Such a decision algorithm would not only effectively solve the problems of rational protein design but also allow for a quite detailed mapping of protein space.

Of course there is no algorithm available that *unambigously* decides whether a structure ψ is the ground state configuration of sequence x. Knowledge-based potentials derived from the data of known protein structures, however, have proved to be a powerful tool for judging whether a given sequence x is a good candidate for folding into the prescribed structure ψ. A serious restriction applies, however: the structure ψ must be a native protein conformation. In this contribution we make use of a set of *knowledge-based potentials of mean force* that have been derived by Sippl and co-workers [27, 46] based on two fundamental assumptions:

(1) Proteins fold into their thermodynamic ground states, i.e., the conformation of a protein minimizes a potential energy function.

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

These potentials are part of the software package `PROSA II` which has been used successfully to identify native folds of proteins among large sets of possible conformations [27, 47, 48]. They hold promise, thus, to be a useful vehicle for exploring protein space. While the `PROSA` potentials have been very successful, a variety of different approaches to knowledge-based, empirical potentials of mean force have been persued some of which are neither based on pair interactions nor on Boltzman's law, see, e.g., [3, 5, 20, 21, 22].

In the present study we use the PROSA potentials to investigate the following questions:

- How easy is inverse folding?
- Is inverse folding still feasible with severly restricted amino acid alphabets?
- How are sequences that fold into a given shape distributed in sequence space?
- Are there neutral networks in protein space?

2. Methods

The starting point for our discussion is a potential function $W(x,\psi)$ evaluating the energy of a sequence x when folded into a shape ψ which is defined by the spatial coordinates of its C^α and C^β 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 [8, 27, 46, 47, 48]. The potential functions used in this program take the form

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

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 Euklidean distance $r = \mathbf{d}_{ij}^\gamma$ of 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 [4, 5, 33]. The parameter χ serves as a (crude) quantitative measure for the surface-exposure of the residue a.

The quality of the fit of the sequence x to structure ψ, as compared to its fit with other structures is conveniently expressed by the *z-score*

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

where $\overline{W}(x)$ is the average energy of sequence x in all conformations in the database and $\sigma_W(x)$ is the standard deviation of the corresponding distribution. Use of the *z-score* instead of just the energy W in the inverse folding introduces a kind of negative design [13, 26] that avoids sequences with multiple stable structures [1].

Figure 2 shows that the z-score correlates well with the rms deviation of alternative structures from the wild type structure [31]. A second line of evidence comes

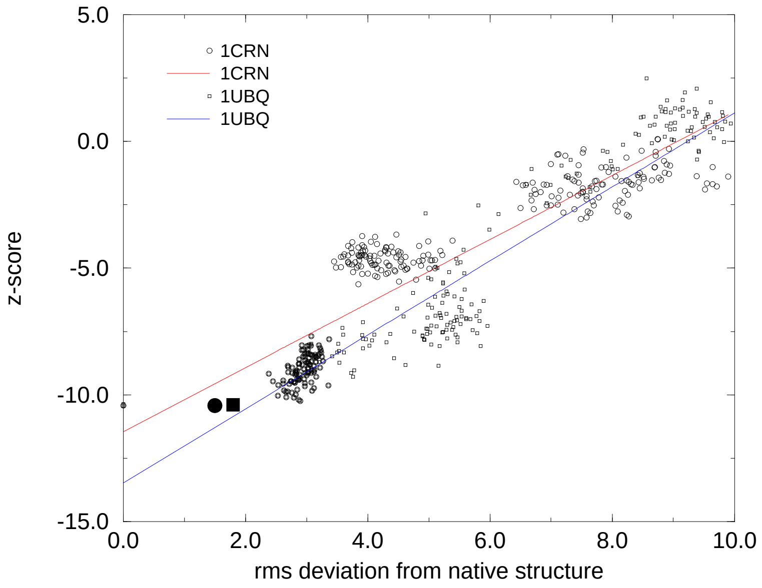

Figure 2: z-scores of perturbed protein structures.
A set of perturbed protein structures was produces by "heating" the structure in a molecular dynamics simulation and then cooling (and compactifying) it again, [31]. The two big symbols show the native structures at their experimental resolutions.

Table 1. z-scores of the sequence of *aspartat aminotransferase* (E.C.2.6.1.1) of *E.coli* computed for four different X-ray structures.

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

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.

from X-ray structures measured at different resolution by different groups (See table 1). Their z-scores indeed improve with increasing resolution of the structure determination.

We may therefore interpret the z-score difference between the wild type sequence $\hat{x}$ and a test sequence x as a measure of distance between their respective native folds. We set

$$D_\psi(x,\hat{x}) \stackrel{\text{def}}{=\!=} \begin{cases} z(x,\psi) - z(\hat{x},\psi) & \text{if } z(x,\psi) > z(\hat{x},\psi) \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

i.e., we accept x as a sequence folding into ψ if its z-score is at least as good as the z-score of the wild type sequence $\hat{x}$. The z-score can thus be used as an approach to *inverse folding*: Given a fixed backbone conformation ψ, we may search for sequences x that give z-scores $z(x,\psi)$ at least as low as the z-score of the native sequence $\hat{x}$. Of course, only native structures ψ that are already in the database can be explored with 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 cost function $D_\psi(x,\hat{x})$, or simply at the minima 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 (exchanges of single amino acid) that are accepted if and only if the cost function (D_ψ) decreases. If no advantageous mutation can be found, the procedure terminates in a local optimum of the cost function. The procedure is equivalent to a Monte-Carlo optimization, as used in [44, 45], Typical examples of adaptive walks are shown in Figure 3.

The length L_{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_{adap}/n indicates the abundance of sequences that fold into the wild type structure ψ, because rare structures can only be reached 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 2 in the following section.

Whether the sequences predicted by our inverse folding procedure do indeed fold into the desired structure can ultimately only be answered by experiment. One way to test the plausibility of a computed sequence is to try to predict its secondary structure and to compare it with 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

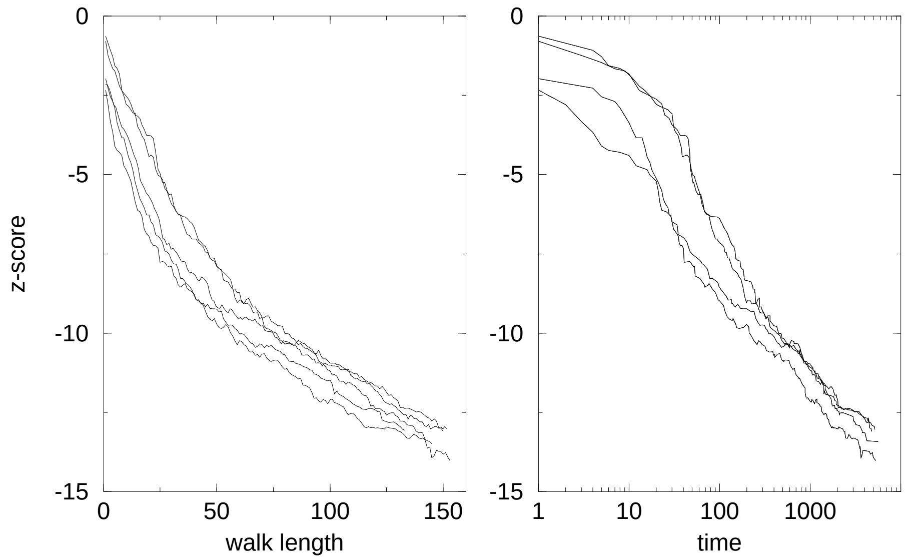

Figure 3: The z-score along adaptive walks on the 2TRX structure is plotted as a function of the walk length (l.h.s) and as a function of the time, i.e, the number of attempted steps (r.h.s). The final sequences derived from the adaptive walks on the 2TRX structure had z-scores about 50% better than the score of the wildtype sequence (-9.22).
This may seem surprising at first. There is no reason, however, why the wild-type z-scores should be optimal; for instance, too much structural stability might even be detrimental to the function of a protein. Alternatively, this effect might be attributed to the limited accuracy of the potentials of mean force. Table 1 shows that the z-scores of the native structures improve with increasing resolution, i.e., that the data contained in the PDB files is by no means completely accurate. Extrapolation of the ubiquitin data in figure 2, for instance, suggest that about six z-score units might be explained by the limited resolution of the experimental structures. We should therefore expect, that sequences with z-scores better than that of the native structures can be found.
Although the scores are steadily improving even after about 150 steps the time, i.e., the number of tested mutants, needed to find an improved sequence, becomes very large towards the end of the simulation.

over 70% for the assignment of residues to helix, strand and loop regions [39]. Since our inverse folded sequences have no or little homology to known sequences we have to expect somewhat lower accuracies.

We mostly used the program `SOPM` by Geourjon and Deleage [19]. It correctly predicts approximately 65% of the residues in the 2TRX wildtype. Similar results were achieved using the `PHD` method of Rost and Sanders [39]. Figure 4 shows the overlap between the 2TRX secondary structure and the `SOPM` prediction for every 5th sequence from the 5 adaptive walks shown in figure 2. The overlap between

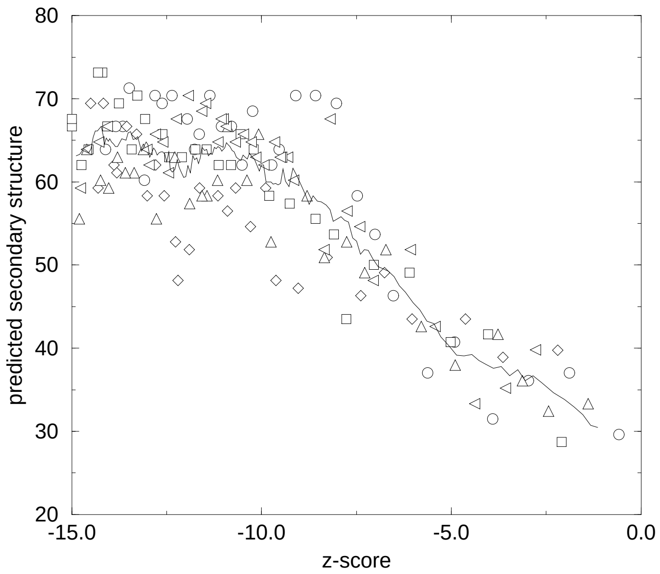

Figure 4: Identity between predicted and 2TRX secondary structure as a function of z-score for sequences from the 5 adaptive walks shown in 3. The full line is a running average of all 5 data sets.

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. The `PROSA` potentials depend only on distances between C^{α} or C^{β} atoms and surface exposure and make no explicit reference at all to secondary structure. The fact that the `SOPM` prediction of the secondary structure of the inverse folded sequences agrees with the wild type secondary structure within the limits of the algorithm's accuracy therefore lends further credibility to the `PROSA` potentials.

3. Results

3.1. 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 the same fold [37]. The distribution of pairwise Hamming distances for 700 sequences with z-score approximately -11 on the 2TRX structure is shown in Figure 5. Although they lie somewhat closer together than random sequences with a typical amino acid composition (taken from the `SwissProt` database), pairs with the maximal Hamming distance $n = 108$ still do occur.

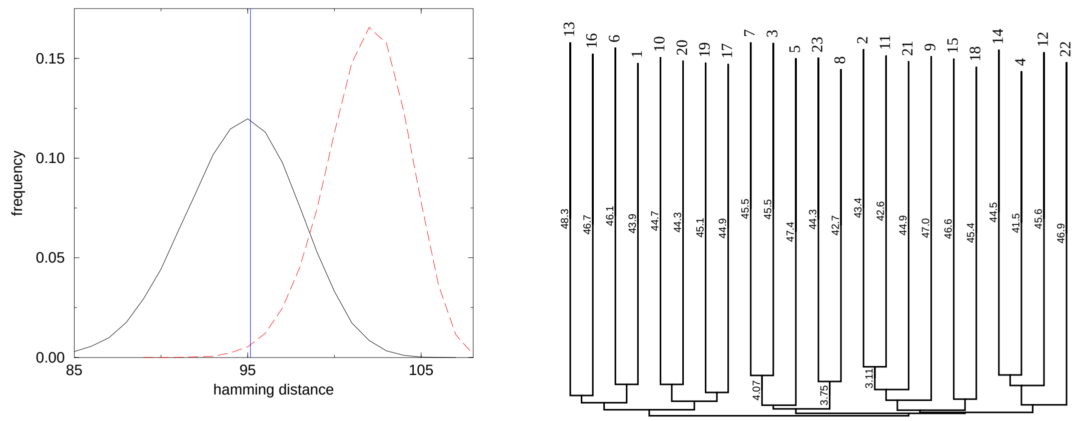

Figure 5: Distribution of pairwise Hamming distances for 700 sequences designed to have z-scores of about -11 on the 2TRX structure (full curve) and for 500 random sequences with typical amino acid composition (l.h.s.). The vertical line at 95.15 is the average distance to the 2TRX sequence. Its position close to the maximum of the distribution indicates that the wild type sequence is not exceptional among the inverse folded sequences. The position of the maximum depends only slightly on the z-score of the inverse folded sequences.
The r.h.s. shows 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.

The distribution of Hamming distances of sequences taken from different adaptive walks have been analyzed in detail for traces of clustering. Tree reconstruction methods, such as neighbor joining, see figure 5, and the split decomposition technique [2] suggest that the sequences with wildtype-like z-scores are distributed essentially randomly in sequence space. The average Hamming distance $\langle d \rangle_{\mathrm{adw}}$ of the end-points of independent walks is about 80 to 90 percent of the chain length. This value is significantly below the expected value of 95% in a 20 letter alphabet. It can be explained, however, by the non-uniform distribution of amino-acids. The $\langle d \rangle$ data are compiled in table 2.

A number of groups have argued that the pattern of hydrophobic versus hydrophilic amino acids (**HP**-pattern) has a dominating influence on protein structure [14, 32, 49]. An example of the **HP**-pattern of inverse folded 2TRX sequences is shown in figure 6.

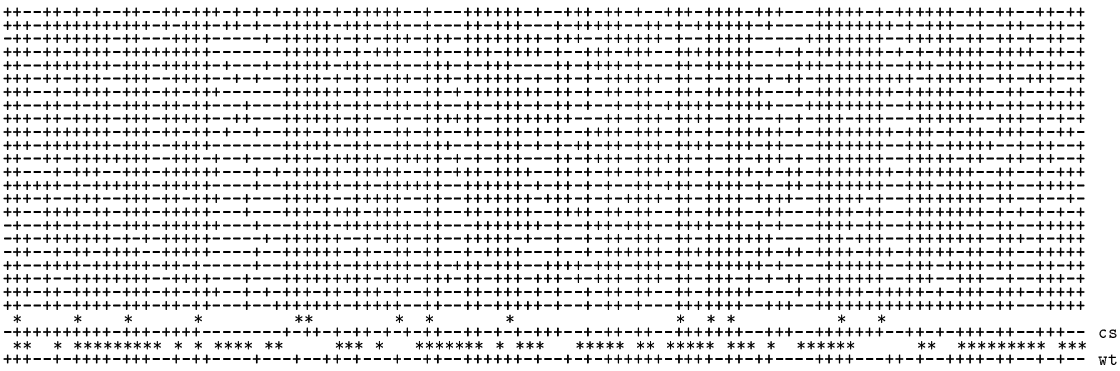

Figure 6: **HP**-patterns of sequences generated by neutral paths, see section 3.2, on the Thioredoxin structure at z-scores comparable to the wild-type. Hydrophilic amino acids (**R**, **N**, **D**, **E**, **Q**, **G**, **H**, **K**, **P**, **S**, **T**, and **Y**) are shown as `+`, hydrophobic amino acids (**A**, **C**, **I**, **L**, **M**, **F**, **W**, and **V**) are represented as `-`. 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 wildtype sequence. The stars between wildtype and consensus sequence highlight the overlap of these two sequences.

There is 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 next section). The pattern shown in figure 6 is typical: while in most positions the type of amino acid conforms the wildtype most of the time, the deviations are substantial. The thioredoxin wildtype sequence contains 72 hydrophilic residues (66.7%), the fraction of hydrophilic residues in figure 6 is 69.7%. We find an average Hamming distance of 33.0 between the patterns in figure 6 while the average distance between random sequences with the same fraction of hydrophilic residues is approximately 45.6. Furthermore there are 14 conserved positions. Taking these into account we would expect a Hamming distance of approximately 40 if the 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.

3.2. The Existence of Neutral Networks in Protein Space

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[42]. A neutral path therefore ends after $\mathcal{L} \leq n$ steps when no mutant of $x_{\mathcal{L}}$ can be found that has Hamming distance $\mathcal{L} + 1$ from the starting point and folds into ψ.

The inverse folded sequences were used as starting points for neutral paths. In our simulations we accept a sequence x_i as folding into the prescribed structure ψ if its z-score, as well as the z-scores computed separately from the C_α and the C_β potentials, is below a threshold value that is at least as low as the wildtype z-score. The substitution frequencies for the production of mutants were computed from the natural frequencies of the amino acids as contained in the `SwissProt` database.

Figure 7 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-scores 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

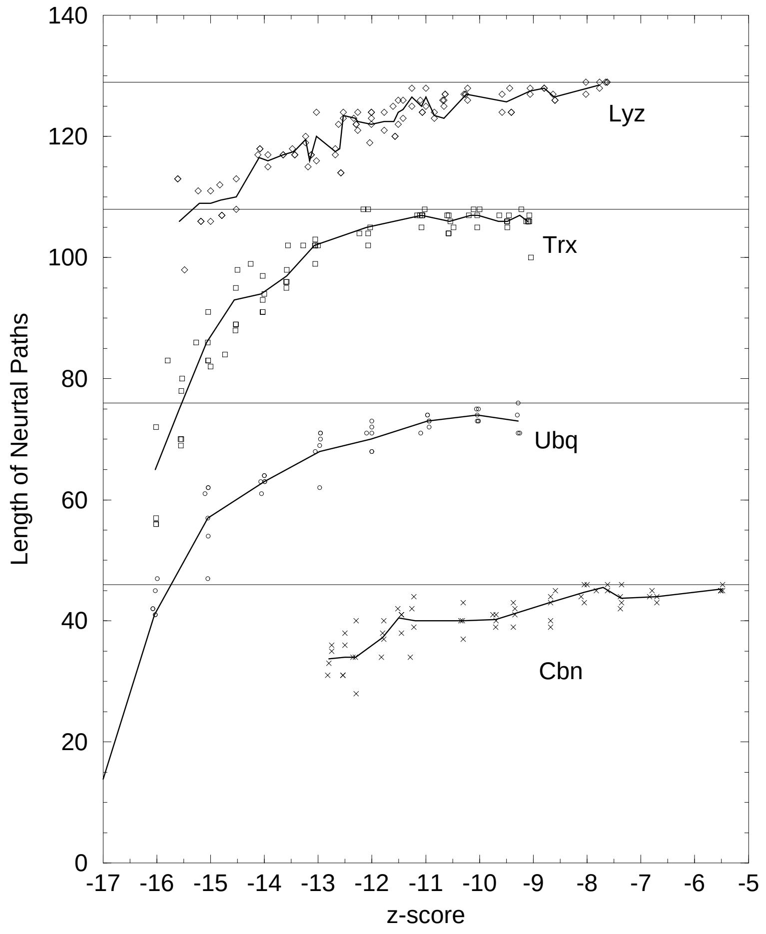

Figure 7: Length of neutral path as a function of the threshold z-score. The solid lines indicate the averages over the available data for each protein. The chain lengths of the four proteins 1cbn, 1ubq, 2trxA, and 1lyz are indicated by thin solid lines for comparison. The rightmost data points for each structure correspond to the wild-type z-scores, see table 2.

paths is still greater than three quarters of the length of the protein. The average values of $\mathcal{L}$ taken over the z-score interval $z_{\text{w.t.}} - 3 \leq z \leq z_{\text{w.t.}}$ are collected in table 2. The average Hamming distances between the endpoints of unrelated neutral path, $\langle d \rangle_{\text{nn}}$, are in the range of 90 to 95 percent 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 end points of neutral paths is somewhat larger than the average distance between the end points of adaptive walks, since a neutral walk has a built-in bias towards sequences that contain a more uniform distribution of aminoacids.

Table 2. Characteristics of Neutral Networks.

Protein	n	z	$\mathcal{L}^*$	L_{adap}	$\langle d\rangle_{\text{adw}}$	$\langle d\rangle_{\text{nn}}$
1cbn	46	−5.50	44.6	17.7	38.7	42.3
1ubq	76	−9.30	72.5	61.9	61.1	66.3
2trxA	108	−9.22	106.3	71.7	87.7	97.5
1lyz	129	−7.70	126.2	58.2	106.0	118.7

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

The vast extension of the network of neutral paths suggest that extensive *neutral networks* of sequences folding into the same structure percolate the entire sequence space [38]. The existence of extensive neutral networks meets a claim raised by Maynard-Smith [35] for protein spaces that are suitable for efficient evolution. The evolutionary implications of neutral networks are explored in detail in [30]. Empirical evidence for a large degree of *functional* neutrality in protein space was presented recently by Wain-Hobson and co-workers [34].

3.3. 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 aminoacids. 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 were used. Surprisingly, however, we observed substantial difference between different alphabets that all contain both hydrophilic and hydrophobic aminoacids.

The two-letter alphabet **AD** gives very poor results, for instance, while other combinations of just one hydrophilic and one hydrophilic amino acid, such as **LS** or **DL**, yield wild-type like z-scores for all four of our test proteins. It is not surprising that **ADL** and **ADLG** yield good sequences since **DL** is already

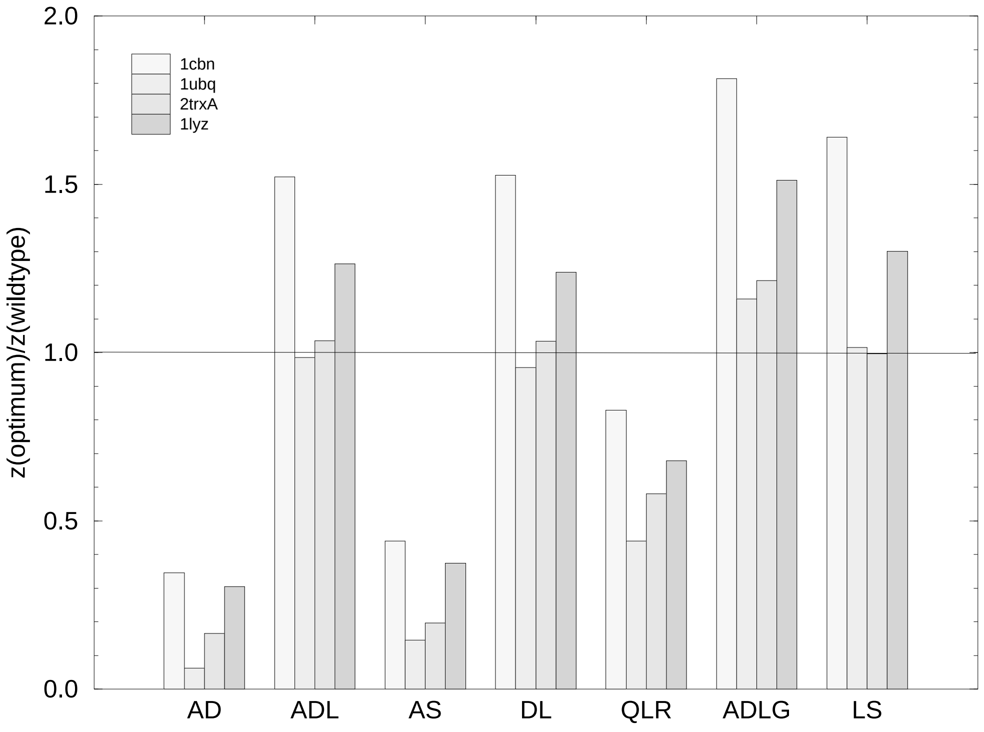

Figure 8: Inverse Folding with restricted alphabets.
The data clearly show that some restricted alphabets allow for inverse folding of native structures while we could not find acceptable z-scores with other alphabets. It is also interesting to note that the quality of best solutions depend on the structure 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.

sufficient. The inverse folded sequences in these alphabets do, however, contain a substatial fraction of **A** and **G**. The alphabet **ADLG** has been proposed as a candidate for a primordial set of aminoacids, before the full genetic code was developed. 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 [11, 41] 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

$$\begin{array}{lcccc} & \text{Cbn} & \text{Ubq} & \text{Trx} & \text{Lyz} \\ \hline \mathcal{L} & 40.2 & 58.0 & 87.5 & 109.0 \\ \mathcal{L}/n & 0.87 & 0.76 & 0.81 & 0.84 \\ \hline \end{array} \tag{4}$$

Note that the distance between random sequence is $0.75n$ in a four letter alphabet. Neutral paths thus extend well beyond the expected distance of random sequences even in the highly restricted **ADLG** alphabet.

The best z-score relative to the wild type have the order,

$$\text{CBN} > \text{LYZ} > \text{TRX} > \text{UBQ}\,, \tag{5}$$

irrespective of the amino acidalphabet. We find the same ordering for $\langle d\rangle_{\text{nn}}/n$ and $\langle d\rangle_{\text{adw}}/n$ while the relative lengths of L_{adap}/n give the reverse order. This suggests that "designability" is indeed an intrinsic property of the structure and that it may vary significantly among wildtype structures [25].

4. Discussion

The approach to inverse folding outlined in this work may be useful as a tool for studying the global properties of the sequence-structure relations in proteins, as well as for protein engineering in biotechnological applications.

We found that sequences with z-scores as good as or better than the wild-type z-score can indeed be found using a simple adaptive walk in sequence space. A more sophisticated optimization technique is not necessary since local minima in the high dimensional sequence space are rare. We found that there is essentially no homology between the inverse folded sequences, the distribution of the amino acids is essentially random. The neutral paths on protein structures extend to the length of the amino acid sequence at z-scores comparable to the wild-type score and better. The existence of very long neutral walks on all native protein structures is a strong indication for the existence of *neutral networks* percolating protein space. The same qualitative results were obtained for some, but not all, restricted amino acid alphabets. In this respect, the sequence-structure map of proteins seems to be very similar to the nucleic acid case.

RNA exhibits *shape space covering*, that is, any common structure can be found within a small radius in sequence space centered around an arbitrary reference point [42, 23, 24]. It is not clear at this point whether proteins (or the knowledge-based potential model of protein space considered here) shares this property. Sander and Schneider [40] have argued that sequences with more than some 30% sequence homology will give rise to the same fold. Random graph models of neutral networks [38], 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.

Acknowledgements

This research was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj.Nos. 10857-CHE and 11601-GEN, and a grant from the *Diversity Biotechnology Consortium*, New Mexico.

References

[1] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.*, 252(4):460–471, 1995.

[2] H. J. Bandelt and A. W. M. Dress. A canonical decomposition theory for metrics on a finite set. *Adv. math.*, 92:47, 1992.

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

[4] J. U. Bowie, N. D. Clarke, and C. O. Pabo. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, 7:257–264, 1990.

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

[6] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins*, 21:187–195, 1995.

[7] C. J. Camacho and D. Thirumalai. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Lett.*, 71:2505–2508, 1993.

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

[9] H. S. Chan and K. A. Dill. Sequence space soup. *J. Chem. Phys.*, 95:3775–3787, 1991.

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

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

[12] T. Defay and F. E. Cohen. Evaluation of current techniques for ab initio protein structure prediction. *Proteins*, 23:431–445, 1995.

[13] W. F. DeGrado, Z. R. Wasserman, and J. D. Lear. Protein design, a minimalist approach. *Science*, 243:622–628, 1989.

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

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

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

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

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

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

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

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

[22] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. *Ismb*, 3:154–161, 1995.

[23] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by

exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996.

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

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

[26] M. H. Hecht, J. S. Richardson, D. C. Richardson, and R. C. Ogden. De novo design, expression, and characterization of felix: a four-helix bundle protein of native-like sequence. *Science*, 249:884–891, 1990.

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

[28] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

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

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

[31] M. Jaritz. Personal Communication, 1995.

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

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

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

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

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

[37] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, 1994.

[38] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 1996. in press, SFI preprint 95-07-058.

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

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

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

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

[43] E. Shakhnovich, V. Abkevich, and O. Ptitsyn. Conserved residues and the mechanism of protein folding. *Nature*, 379:96–98, 1996.

[44] E. I. Shakhnovich and A. M. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci*, 90(15):7195–7199, 1993.

[45] E. I. Shaknovich and A. M. Gutin. A new approach to the design of stable proteins. *Protein Engineering*, 6(8):793–800, 1993.

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

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

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

[49] S. Sun, R. Brem, H. S. Chan, and K. A. Dill. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, 8:1205–1213, 1995.

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

[51] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.*, 1996. in press, Santa Fe Institute Preprint 96-04-016.

[52] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bull. Math. Biol.*, 55:1183–1198, 1993.

[53] K. Yue and K. A. Dill. Sequence structure relationship of proteins and copolymers. *Phys. Rev. E*, 48:2267–2278, 1993.

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

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