R. Giegerich et al. (eds.)
Computer Science and Biology. GCB'99 Hannover, Germany. Univ. Bielefeld 1999, pp. 107-112

AN EFFICIENT POTENTIAL FOR PROTEIN SEQUENCE DESIGN

G. Weberndorfer[a,*], I.L. Hofacker[a], P.F. Stadler[a,b]

[a]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52737, 52793 (fax) E-Mail:
{gw,ivo,studla}@tbi.univie.ac.at
*Address for correspondence
[b]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA

1. Inverse Folding

The native structure ψ of a given amino acid sequence x corresponds to the minimum of its free energy, $W(x, \psi)$. If this energy function W were known, the native fold could in principle be predicted from the amino acid sequence by energy minimization in conformation space. Although the energy function is complex and the computational problems are formidable, this is in principle a straightforward recipe. It has indeed been used successfully to investigate the sequence-structure relation for RNA molecules [10].

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 (ground state) of *each* sequence is equal to 0. This, of course, amounts to solving the protein folding problem for each possible sequence first.

Suppose we are given a knowledge-based potential function $W(x, \psi)$ evaluating the energy of a sequence x when folded into a structure ψ. The structure (or fold) ψ is customarily defined by the spatial coordinates of a subset of its atoms, usually the C^α- and/or C^β-atoms. A series of studies using a variety of different knowledge based potentials by different groups 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 values W for x in its conformation space, see e.g. [8, 4]. This allows the construction of an energy scale (*z-score* [5]) by which conformations of different sequences can be compared:

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

Here the average energy $\overline{W}(x)$ of sequence x is obtained from evaluating x in all conformations in a database, $\sigma_W(x)$ is the standard deviation of the corresponding distribution. Empirically, native folds have z-scores in a narrow characteristic range [13].

It is therefore reasonable to assume that x folds into ψ if the $z(x,\psi)$ lies in the native range. Under these premises, knowledge-based potentials can be used to solve the *inverse folding problem* of finding a sequence that folds into a predefined conformation. Indeed, inverse folding translates into optimizing the z-score with fixed ψ. Using inverse folding techniques, large scale statistical surveys of the sequence-structure map of poly-peptides can be conducted without actually solving the folding problem [2, 1].

2. Delauney Tessellation of Protein Structures

Protein potentials based on Delauney tessellations were originally proposed by A. Tropsha *et al.* [11, 15, 9]. The amino acids in a protein chain are represented by their C^α-atoms or C^β-atoms. Then the Delauney tessellation of the resulting point set is computed using `qhull` [3]. It defines nearest neighbors in an unambiguous and parameter-free way by decomposing the structure into tightly packed tetrahedra which have the C^α-atoms as their corners. From the database of residue frequencies in these tetrahedra, a potential is extracted using the *inverse Boltzmann law* as described in [12].

Denote the observed frequency of a contact tetrahedron consisting of the amino acids i, j, k, l by f_{ijkl} and let p_{ijkl} be the *a priori* expected frequency of such a tetrahedron, which is proportional to the product of the single amino acid frequencies. The energy contribution of the contact tetrahedron is then given by the log-likelihood ratio

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}\,. \tag{2}$$

The contact energy $W_C(x,\psi)$ of sequence x when threaded onto structure ψ is then the sum of the energy contributions of all tetrahedra into which the structure is decomposed.

This approach suffers from a significant shortcoming: The Delauney tessellation by itself does not distinguish between surface residues and buried amino acids. However, it is straight forward to identify those triangles (faces of the tetrahedra) that are exposed to the outside of protein. A buried triangle will appear in two tetrahedra, while a surface triangle appears only once.

It is now straight forward to introduce a surface term based on the log likelihood ratios q^s_{ijk} of triples i, j, k of amino acids in surface triangles.

3. Results

The computation of the Delauney tessellations of a protein structure can be quite costly. While this may be a problem for threading applications, for inverse folding the tessellations of the target structure as well as the collection of reference structures needed for computing the z-score have to be computed only once and can be kept in memory. The tessellations can then be used for arbitrary many z-score evaluations, as long as the structure is kept constant. The tessellation potential is therefore optimally suited for large scale explorations of protein space using inverse folding techniques.

Table 1. z-scores of native and misfolded structures
For five pairs of proteins with the same sequence length z-scores were calculated for the native sequence on its respective structure and with sequence and structure swapped. A comparison of z-scores shows, that each sequence is able to identify its native structure and, combined potentials increase the discriminating power. The last columns show `PROSA` scores for comparison (note good `PROSA` scores are negative).

Sequence		Z - SCORES					
		Combined Pot.		Contact Pot.		PROSA	
PDB id.	l_{seq}	A	B	A	B	A	B
A: `1cbh`	36	7.81	-0.98	7.78	-0.44	-4.76	-1.12
B: `1ppt`		-2.68	1.67	-1.81	1.03	0.13	-2.79
A: `1fdx`	54	5.65	-0.38	4.76	-0.00	-7.46	-1.29
B: `5rxn`		-0.07	2.53	-0.23	2.92	-1.38	-5.87
A: `1ubq`	76	8.81	-3.47	8.73	-2.38	-9.25	0.30
B: `4icb`		-1.25	6.32	-1.29	5.26	-1.05	-10.06
A: `2trx`	108	9.28	-1.36	8.48	-1.92	-7.04	-0.61
B: `1rro`		-2.43	6.08	-2.20	4.87	0.13	-10.88

Any useful potential must be able to discriminate between native and misfolded structures. The simplest test uses unrelated structures of equal length and checks whether the potential correctly assigns each sequence to its native structure. The tessellation potential passes this test, Table 1.

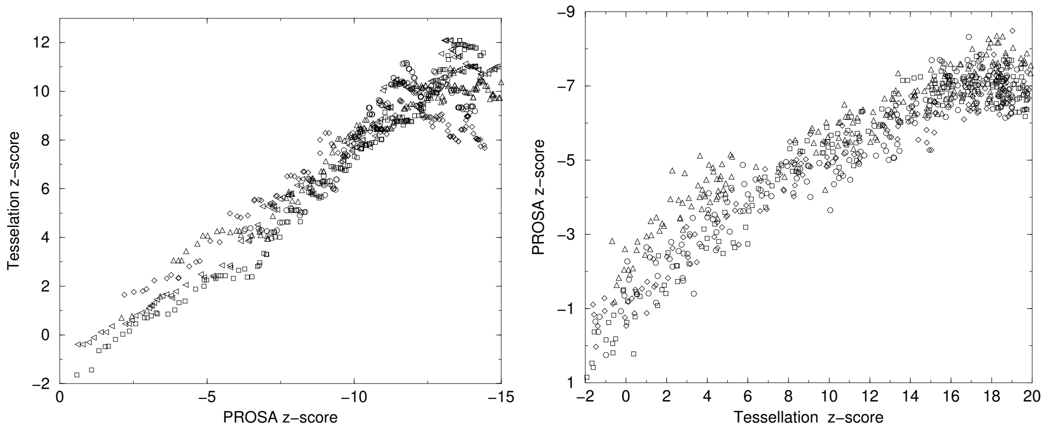


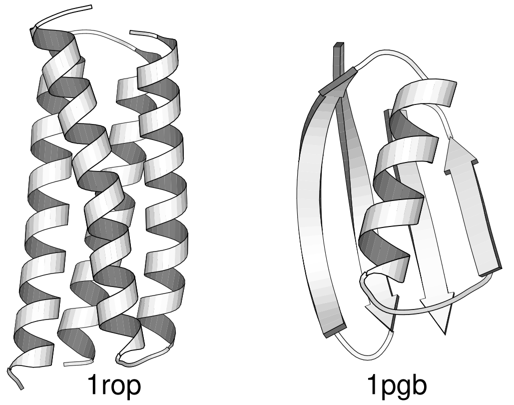
Figure 1. Comparison of z-scores from `PROSA` and tessellation potential.
Sequences are sampled by adaptive walks in one potential (x-axis) and re-evaluated in the other one (y-axis). Data are for the thioredoxin structure 2trx.

A more stringent test can be constructed by comparison with other knowledge based potentials such as Sippl's `PROSA II` [13], or the neural network (`NN`) of Lapedes *et al.* [7]. Sequences produced from inverse folding in one potential are re-evaluated in another potential. Ideally, z-scores in both potentials should correlate well with a slope close to unity. This test ensures that the potentials do not produce false positives, see Figure 1.

We found a good correlation between the tessellation potential and e.g. PROSA. However, usually sequences optimized in one potential exhibit significantly smaller z-scores in another. This reflects the inaccuracies and inconsistencies between different potentials. Figure 1 indicates that sequences should therefore be optimized until a z-score several units better than the wild type level is reached. In addition, Zhang and Skolnick [14] recently estimated that the z-score of a protein of 100 residues should be better than 15, significantly larger than the scores of native sequence-structure pairs obtained from currently available potentials.

4. JANUS as an Example

In a quite spectacular experiment Dalal *et al.* [6] have designed a protein, Janus that has 50% sequence identity with a predominantly β-sheet protein (1pgb), but which adopts a four-helix bundle conformation (the structure of the dimer of 1rop) and possesses the attributes of a native protein. Starting from the 1pgb sequence they mutated selected residues to adapt the sequence to the Rop fold, usually by replacing them with the corresponding amino acid from Rop (only 9 positions in Janus differ from both Rop and 1pgb).




```
MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARF   (1rop)
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE   (1pgb)
MTKKAILALNTAKFLRTQAAVLAAKLEKLGAQEANDNAVDLEDTADDLYKTLLVLA   (Janus)
```

Figure 2. 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.* [6].

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 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. We therefore use

Table 2. z-score evaluation of wildtype and `janus` sequences

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

the structure of the Rop dimer in the experiment, making sure that both copies of the sequence forming the Rop dimer are always mutated together. The mutation is accepted if the z-score of the mutated sequence is no worse than the wildtype z-score of `1rop`. The experiment is terminated when no further acceptable mutations can be introduced.

Our simulations easily produced sequences with more than 50% sequence identity with the `1pgb` target sequence. The final sequences of 10 independent experiments showed an average sequence identity of approximately 87% while retaining wild-type like z-scores for the `1rop` target structure. Table 2 shows the z-scores of the two wildtype sequences, the Janus sequence of Dala *et.al.*, and the average for the 10 simulated sequences.

Acknowledgements. This work was supported in part by the *Jubiläumsfond der Österreichischen Nationalbank* Proj. 6792.

References

[1] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler. Exploring protein sequence space using knowledge based potentials. *J. Comp. Biol.*, 1999. submitted, Santa Fe Institute preprint 98-11-103.

[2] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.

[3] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Software*, 22:469–483, 1996.

[4] 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.

[5] G. Casari and M. J. Sippl. 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, 1992.

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

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

[8] 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.

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

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

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

[12] 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.

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

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

[15] W. Zheng, S. J. Cho, I.I.Vaisman, and A. Tropsha. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In R. B. Altman and al., editors, *Pacific Symposium on Biocomputing'97*, pages 487–496, Singapore, 1997. World Scientific.