New Techniques for the Construction of Residue Potentials for Protein Folding

Arnold Neumaier, Stefan Dallwig, Waltraud Huyer, and Hermann Schichl *

Institut für Mathematik, Universität Wien
Strudlhofgasse 4, A-1090 Wien, Austria

Abstract. A smooth empirical potential is constructed for use in off-lattice protein folding studies. Our potential is a function of the amino acid labels and of the distances between the C_α atoms of a protein. The potential is a sum of smooth surface potential terms that model solvent interactions and of pair potentials that are functions of a distance, with a smooth cutoff at 12 Ångström. Techniques include the use of a fully automatic and reliable estimator for smooth densities, of cluster analysis to group together amino acid pairs with similar distance distributions, and of quadratic programming to find appropriate weights with which the various terms enter the total potential. For nine small test proteins, the new potential has local minima within 1.3–4.7Å of the PDB geometry, with one exception that has an error of 8.5Å.

Keywords. protein folding, tertiary structure, potential energy surface, global optimization, empirical potential, residue potential, surface potential, parameter estimation, density estimation, cluster analysis, quadratic programming

1991 MSC Classification. primary 92C40; secondary 62H30, 90C20

1 Overview

The protein folding problem is the task of understanding and predicting how the information coded in the amino acid sequence of proteins at the time of their formation translates into the 3-dimensional structure of the biologically active protein. A thorough recent survey of the problems involved from a mathematical point of view is given by Neumaier [22].

* The authors gratefully acknowledge support of this research by the Austrian Fond zur Förderung der wissenschaftlichen Forschung (FWF) under grant P11516-MAT.

The forces in a protein molecule are modeled by the gradient of the potential energy $V(s,x)$ in dependence on a vector s encoding the amino acid sequence of the molecule and a vector x containing the Cartesian coordinates of all essential atoms of a molecule. In an equilibrium state x, the forces $\nabla V(s,x)$ vanish, so x is stationary; and for stability reasons we must have a local minimizer. The most stable equilibrium state of a molecule is usually the native (tertiary) state. Thus, finding the native state of a protein molecule with sequence s is considered to be more or less equivalent to finding the global minimizer $\hat{x}$ of $V(s,x)$.

Therefore, modeling a protein molecule amounts to deciding on the atoms considered to be essential and to specifying the contribution of the various interactions to the potential. Since the work to find the global minimizer increases drastically (and possibly exponentially) with the dimension of x, it is customary to use for larger proteins a reduced description that treats only very few atoms in each amino acid as essential.

LEVITT & WARSHEL [17,18] were the first to show that reduced representations may work; they used C_α atoms and virtual atoms at side chain centroids. OOBATAKE & CRIPPEN [24] simplified further by only considering the C_α atoms. This is sufficient since there are reasonably reliable methods (HOLM & SANDER [11,12]) that compute a full atom geometry from the geometry of the C_α atoms. (All atom representations are used as well, but limited to the prediction of tiny systems such as enkephalin.)

Our potential is a sum of smooth *surface potentials* that model amino acid-solvent interactions and of smooth *pair potentials* that model amino acid-amino acid interactions. As in [24], we take as essential only the C_α atoms.

Traditionally, pair potentials are determined by assuming that a set of known structures, generally taken from the Brookhaven Protein Data Bank (PDB) [3,35,30], is an equilibrium ensemble of structures, so that, up to a constant factor determined by the temperature, the energy can be calculated from Boltzmann's law and statistics on the known structures. In order to obtain useful statistics, the protein structures used must be carefully selected; see, e.g., HOBOHM et al. [9]. A more detailed overview can be found in SIPPL [28]. Other empirical potential construction techniques are discussed in BAUER & BEYER [1] and ULRICH et al. [34]. The fact that the potential is directly derived from geometric data implies that it automatically takes account of solvation and entropy corrections; on the other hand, one only gets a mean potential of low resolution. Reconstructions using mean potentials are reported by SUN [31] for apamin (18 residues) and mellitin (26 residues) using genetic algorithms, by SUN [32] for mellitin, APPI (36 residues) and crambin (46 residues) using simulated annealing, by GUNN et al. [6] for myoglobin (153 residues) using a combination of simulated annealing and genetic algorithms, and by SIPPL et al. [29] for myoglobin and lysozyme (129 residues) by an assembly process using a fragment database. In all these papers, results are only 'native-like' when compared with the experimental structures.

To find appropriate empirical pair potentials from the known protein structures in the Brookhaven Protein Data Bank, it is necessary to calculate densities for the distance distribution of C_α-atoms at given bond distance d and given residue assignments a_1, a_2. Up to a constant factor that is immaterial for subsequent structure determination by global optimization, the potentials then emerge as the negative logarithm of the densities. Since a huge number of pair potentials is required, fully automatic and reliable density estimators are necessary.

The only density estimators discussed in the protein literature are histogram estimates. However, these are nonsmooth and thus not suitable for global optimization techniques that combine local and global search. Moreover, histogram estimates have, even for an optimally chosen bin size, the extremely poor accuracy of $O(n^{-1/10})$ only, for a sample of size n. The theoretically attainable accuracy of the densities is much better, namely $O(n^{-1/2})$. (See HALL & MARRON [7] and the survey by JONES et al. [14].)

We therefore use smooth density estimation techniques that are more reliable than the histogram estimates. To improve the reliability for rare amino acid pairs, we use clustering techniques that identify 'similar' pairs that can be modeled by the same density.

Since the precise form of the potential and the estimation procedure are still in flux, we only discuss the main features but not all the details.

2 Empirical Residue Potentials

We assume that proteins are coded by a sequence

$$s = (s_1, \ldots, s_n)$$

of n amino acids labelled by $s_i \in \{1, \ldots, 20\}$ in one-to-one correspondence with the names of the 20 natural amino acids, and that

$$x_s = (x_1, \ldots, x_n)$$

is a vector listing the Cartesian coordinates $x_i \in \mathbb{R}^3$ of the C_α atoms in the native geometry of a protein molecule with sequence s.

An ideal empirical potential function on the residue level is a function V that assigns to each sequence-coordinates pair (s, x) an *energy* $V(s, x)$ such that

$$x_s \in \mathrm{argmin}_x \quad V(s, x), \tag{1}$$

i.e., x_s is a solution of the global optimization problem associated with the sequence s.

In practice, we have a finite database of known pairs with limited accuracy, and we want to satisfy (1), at least approximately, for the pairs in the data base. To make best use of current optimization technology, it is desirable to have a smooth (i.e., twice continuously differentiable with respect to x) potential. This allows for robust local optimization (e.g., [5,25]), and can be combined with global search techniques such as simulated annealing (e.g., [15,32]), genetic algorithms (e.g., [10,31]), smoothing methods (e.g., [21,16]) or branch and bound techniques (e.g., [20]) to approach the global minimizer for sequences s with unknown native geometry.

Unfortunately, the approach of determining empirical potentials from equilibrium data is intrinsically limited, even if we assume complete knowledge of all equilibrium geometries and their energies. It is obvious that statistical potentials cannot define an energy scale, since multiplication of a potential by a positive, constant factor does not alter its global minimizers. But for the purpose of tertiary structure prediction by global optimization, this does not not matter.

A more serious limitation comes from a nonuniqueness theorem (stated and proved in NEUMAIER [23]) that shows that the set of local and global minimizers and stationary points of any family of potentials $V(s,x)$ can be reproduced by an infinite family of other potentials. Thus, no set of equilibrium geometries can determine the true effective potential energy function. This complements the findings of THOMAS & DILL [33] (through a simulation study) and BRYNGELSON [4] (by theoretical arguments) that statistical potentials do not quantitatively reflect the true energies. In particular, empirical potentials solely derived from databases of equilibrium data will probably never be useful for dynamical studies. Moreover, the "energies" computed by an empirical potential may have little to do with real energies, as has been recently pointed out by BEN-NAIM [2].

While this is disappointing, the nonuniqueness theorem also shows that if *some* empirical potential is able to predict correct protein folds then many other empirical potentials will do so, too. Thus, the construction of empirical potentials for fold prediction is much less constrained than one might think initially, and one is justified in using additional qualitative theoretical assumptions in the derivation of an appropriate empirical potential function.

Our potential,

$$V(s,x) = \sum_a z_a V_a(s,x) + \sum_\gamma z_\gamma V_\gamma(s,x), \qquad (2)$$

is a weighted sum of smooth *surface potentials* $V_a(s,x)$ that model the total contribution of the interaction of all amino acids with label a with the solvent, and of smooth *pair potentials* $V_\gamma(s,x)$ that model the total contribution of the interactions between all pairs of amino acids classified to be of the same class γ. The constants z_a and z_γ are positive weights scaling these contributions in a manner adapted to the available data.

(While smooth pair potentials are the rule in the literature, surface terms have traditionally been discontinuous; the only potential using smooth surface terms seems to appear in LUND et al. [19], where the surface term is a function of a smooth approximation to the number of neighbors of a C_α atom.)

The class of a pair of amino acids in positions i, k of a sequence s depends on the labels s_i and s_k of the amino acids and the *residue distance* $i-k$, and is specified through a suitably constructed class table, and

$$V_\gamma(s,x) = \sum_{(s_i,s_k,i-k)\text{ of class }\gamma} U_\gamma(r_{ik}),$$

where $r_{ik} = \|x_i - x_k\|$ is the Euclidean distance between the positions x_i and x_k of two C_α atoms, and the *pair potentials* $U_\gamma(r)$ are smooth functions of the distance. Instead of the more customary Lennard-Jones like potentials we use pair potentials that are constant for $r \geq r_{\max}$, with a cutoff at $r_{\max}$ at 12 Ångström, since amino acids at most three residues apart are at distance < 12Å in known proteins. All pair potentials are normalized such that their global minimum has the value 0.

To make the pair potentials flexible but fast to evaluate we chose the $U_\gamma(r_{ik}) = W_\gamma(q_{ik})$ as low degree polynomials in

$$q_{ik} = \left(\frac{\max(c_{\max} - c_{ik}, 0)}{2c_{\max} - c_{ik}}\right)^3,$$

with vanishing constant coefficient, where $c_{\max} = \frac{1}{2}r_{\max}^2$ and

$$c_{ik} = \frac{1}{2}\|x_i - x_k\|^2 = \frac{1}{2}(\langle x_i, x_i\rangle + \langle x_k, x_k\rangle) - \langle x_i, x_k\rangle$$

is the *half squared distance* between two C_α atoms. q_{ik} is bounded in [0,0.125] and smooth, has compact support $[0, r_{\max}]$, and can be evaluated cheaply from inner products, without taking square roots. The coefficients of the polynomials $W_\gamma(q)$ are found adaptively to fit empirical potentials derived from the data; see below.

For similar reasons we chose the surface potentials to be

$$V_a(s,x) = \sum_{s_i=a} W_a(q_i),$$

with low degree polynomials $W_a(q_i)$ in some expression q_i that is larger on the surface than in the interior. Since q_{ik} is large only when the ith and the kth residue are close, and since there are fewer residues close to a residue at the surface, the expression

$$q_i = \frac{1.41}{\sum_{k \neq i} q_{ik}} - \frac{2.65}{n^{1/3}}$$

was used, which has this property. The constants were chosen such that q_i has approximately mean 1 and is reasonably independent of the protein size n. q_i (and hence each surface potential) is a smooth function of the coordinates, except when some position x_k ($k \neq i$) coincides with x_i; then q_i is infinite. Thus, by enforcing $W_a(q) \to \infty$ as $q \to \infty$, the surface potentials also serve the task to keep the amino acids apart.

Given the pair and surface potentials, the weights are then constructed by solving the convex bound constrained quadratic program

$$\begin{array}{ll} \min & \sum_{\text{proteins } s} \|\nabla V(s, x_s)\|^2 \\ \text{s.t.} & \text{all } z_a \geq 1, \quad \text{all } z_\gamma \geq 1. \end{array}$$

The objective function is a nonnegative, convex quadratic in z. It should vanish exactly for an ideal potential; hence minimizing the objective can be expected to give a good approximation to the best potential. The constraints are inspired by the independence assumption of Sippl [28], which amounts to the particular choice

$$\text{all } z_a = 1, \quad \text{all } z_\gamma = 1;$$

our constraint relaxes this assumption in a natural way. It turned out that in the solution, $z_a > 1$ for 9 of the 20 amino acids, and $z_\gamma > 1$ for 15 of the 305 pair classes γ used. (For quadratic programming in general, see, e.g., [5].)

3 Density Estimation

The polynomials $W_a(q)$ $(q = q_{ik})$ and $W_\gamma(q)$ $(q = q_i)$ needed to specify the pair and surface potentials are constructed from the

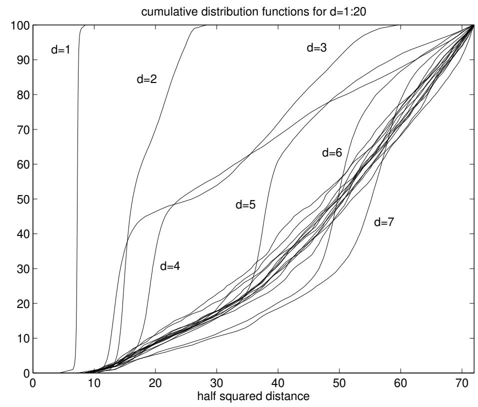

Fig. 1. Cumulative distribution function of half squared distances of amino acid pairs at residue distance d ($d = 1, \ldots, 20$), truncated at $c_{\max} = 72$ (12Åcutoff)

set of such q realized in a data base of 266 proteins with a total of 46100 residues by means of density estimation techniques.

Boltzmann's classical formula

$$\langle f(q) \rangle = \frac{1}{Z} \int e^{-\beta W(q)} f(q) dq$$

for the expectation of functions of a random variable q can be used to justify the use of some multiple $W(q)$ of the negative logarithm of the density $\rho(q) = Z^{-1} e^{-\beta W(q)}$ of q as a useful definition of a potential contribution involving any interesting function q of the coordinates. (For the case when q is an atomic distance, the pros and cons of this recipe are discussed in more detail by Sippl [28].)

For the robust estimation of the pair potentials, some obstacles had to be overcome. There are a huge number of different triples $(s_i, s_k, i - k)$, and to find densities, we needed a way to group them in a natural way together into suitable classes. A look at the cumulative distribution functions (cdf's) of the half squared distances c_{ik} at residue distance $d = i - k$ (w.l.o.g. > 0), displayed in Figure 1, shows that the residue distances 8 and higher behave very similarly; so in a first step we truncated all residue distances larger than 8 to 8.

This left $20 \times 20 \times 8 = 3200$ classes, with some classes being very sparsely populated. For such classes, the error term $O(n^{-1/2})$

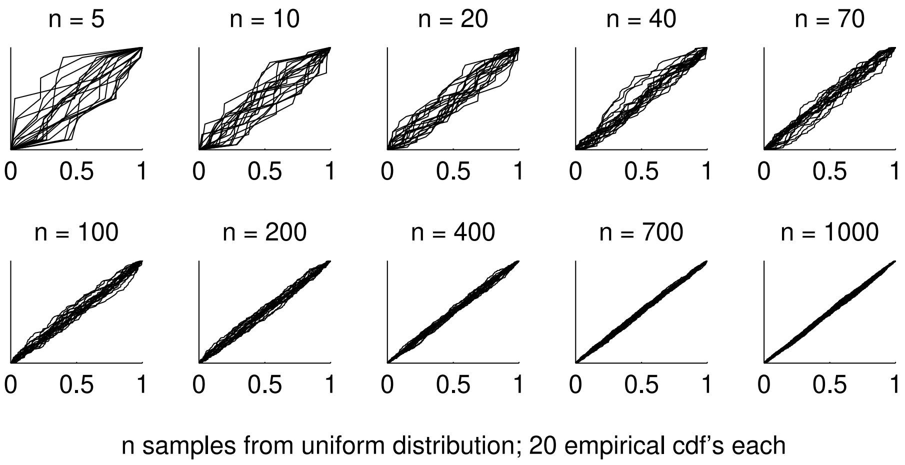

Fig. 2. Empirical cdf's for many samples of increasing size from a uniform distribution

is unacceptably large, and density estimators are intrinsically unstable under variations of small samples. A Monte Carlo test with samples from a uniform distribution displayed in Figure 2 shows that a sample size of at least about 100 is needed to reproduce a cdf and hence a density with a reasonable accuracy.

Therefore it was essential to have an additional mechanism that groups together data for 'similar' amino acid pairs. We performed this task by means of a weighted mean square cluster analysis. Our clustering procedure used linearly interpolated empirical cdf's of the initial classes together with a statistical estimate of their accuracy. We then repeatedly joined the smallest remaining class to the class with the most compatible cdf and recomputed the cdf of the merged class until all classes contained at least 100 sample points. Additional correction phases allowed outliers in some class to migrate to a more suitable class. Applied to the 3200 initial classes, this produced a list of 305 classes with significantly different distributions. (Different clustering procedures are used in some recent residue potential constructions: Ulrich et al. [34] use clustering to group the 210 unordered amino acid pairs of residue distance 3 into 36 classes based on some energy mea-

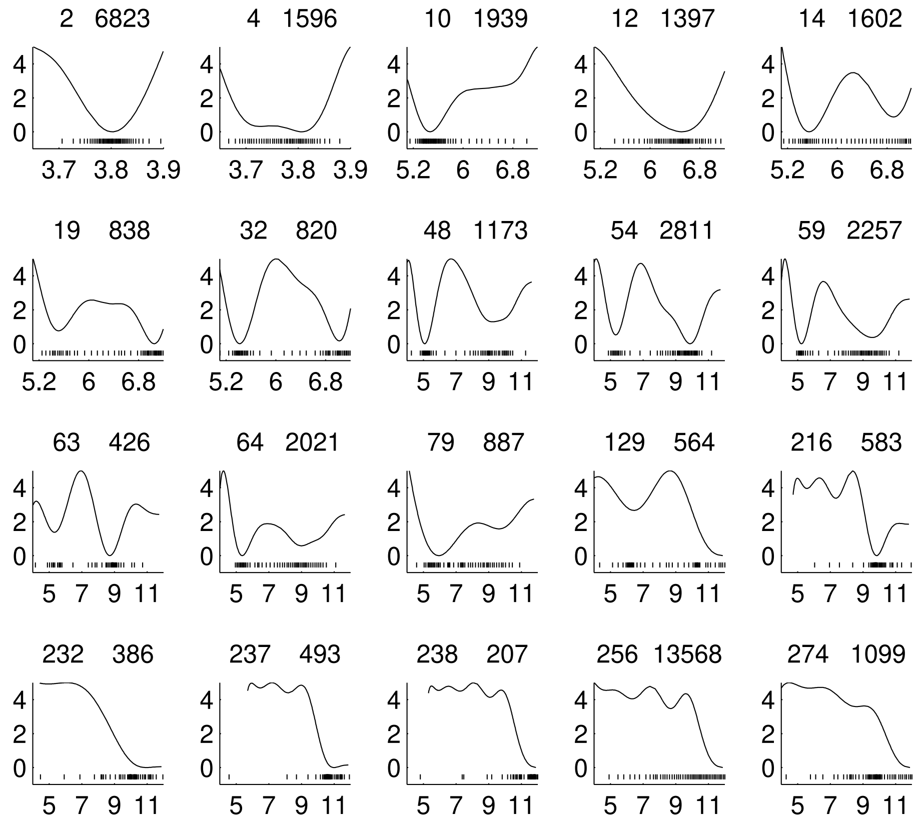

Fig. 3. Pair potentials $U_\gamma(r)$, scaled to move their interesting range to [0,5]. The numbers above each potential denote the class label γ and the number of data points available for the fit. (For example, class 63 gives distance 3 potentials for the amino acid pairs Lys-Asp, Arg-Lys and Glu-Tyr.) The spectrum below each potential consists of 50 lines picked uniformly from the data.

sure, and HUBER & TORDA [13] apply clustering to estimated parameters.)

To compute densities for such a large number of distributions, reliable and fully automatic density estimators are necessary. The only density estimators discussed in the protein literature are histogram estimates. However, these are nonsmooth and thus not suitable for global optimization techniques that combine local and global search. Moreover, for a sample of size n and an optimally chosen bin size, histogram estimates have an accuracy

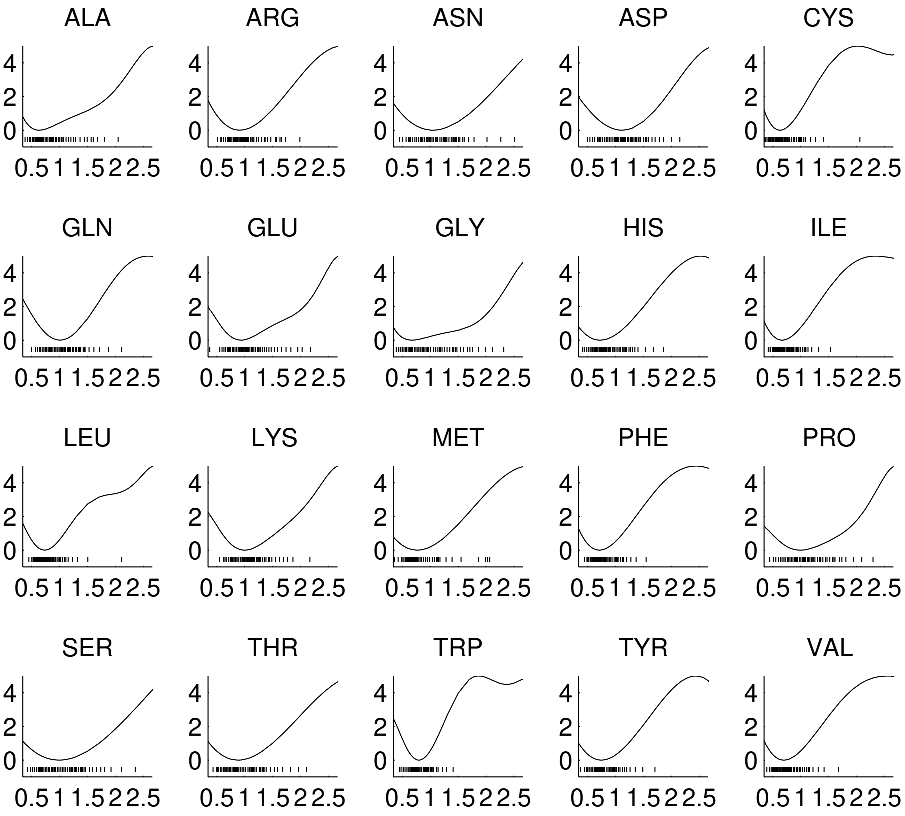

Fig. 4. Surface potentials $W(q)$

of $O(n^{-1/10})$. This is an extremely poor accuracy, far away from the theoretically attainable accuracy $O(n^{-1/2})$ of other density estimators. (To reach $n^{-1/10} = 0.1$ one needs $n = 10^{10}$, while $n^{-1/2} = 0.1$ holds already for $n = 100$.)

One therefore needs a smooth density estimation techniques that is more reliable than the histogram estimates. The automatic estimation poses additional problems in that the traditional statistical techniques for estimating densities usually require the interactive selection of some smoothing parameter (such as the bin size). Some publicly available density estimators are available[1], but these tended to oversmooth the densities. So we tried a num-

[1] see, e.g., `http://solon.cma.univie.ac.at/~neum/stat.html#density`

ber of ideas based on numerical differentiation of the empirical cdf to devise a better density estimator.

The best recipe we found so far is based on the statistical model

$$W(q_l) \approx w_l := \log(q_{l+1} - q_{l-1}) + \log(n/1.5),$$

where q_l is the lth sample point in increasing order. $W(q)$ is estimated from this relation by polynomial regression, and the Bayes criterion of SCHWARZ [26] is used to select the correct polynomial degree.

We tested our recipe on many trial densities by Monte Carlo simulation, e.g., on the normal mixture target densities of JONES et al. [14]. Examples of pair potentials $U_\gamma(r) = W_\gamma(q(r))$ reconstructed in this way are given in Figure 3.

We believe that this recipe gives a reasonably reliable density estimator. But is not yet perfect in that it suffers occasionally from picking an unsuitable polynomial order for the potential, and it tends to give small spurious oscillations at small densities when the original density is not of the form $e^{-W(q)}$ for some low degree polynomial $W(q)$. Unfortunately, these translate into significant oscillations for large values of the corresponding potentials; see Fig. 3 for classes 216, 237, 238, 256. (We are working on replacing polynomials by splines that should suppress these oscillations.)

For the surface potentials, sufficiently many data were available, and no further problems appeared. The resulting potentials are shown in Figure 4. The hydrophobic amino acids are easily recognized as those for which the potential well is at small values of q.

4 Results and Future Work

We tested our new potential by applying a local optimization procedure to the potential of some proteins, starting with the native structure as given in the Brookhaven Protein Data Bank, and observing how far the coordinates moved through local optimization. For a good potential, one expects the optimizer to be close to the native structure. As in ULRICH et al. [34], we measure the distance between optimizer B and native structure A by the distance matrix error

$$DME = \sqrt{\binom{n}{2}^{-1} \sum_{i<k} (r_{ik}^B - r_{ik}^A)^2};$$

it is usually a little larger than the root mean square (RMS) error that is based on optimal superposition.

The results of the optimization for 9 small test proteins, both for the potential with constant weights 1 and with the optimized weights, are given in Table 1. The optimized weights lead to smaller errors; the resulting potentials have minima within 1.3–4.7Å of the PDB geometry, with one exception that has an error of 8.5Å.

Table 1. Distance matrix errors DME (in Å) between optimizers and native structures

PDB code	$z = 1$	optimized z
`1cti`	2.2	4.7
`1gcn`	0.6	1.3
`1mhu`	3.4	3.0
`1mrb`	3.2	2.4
`1mrt`	3.3	2.3
`2eti`	11.8	8.5
`2mhu`	2.3	2.5
`2mrt`	4.0	2.4
`3znf`	5.5	2.3

At present, the data base used for the fit was not specially selected to avoid homologous proteins. Thus, a further improvement can be expected from using data for one of the specially prepared lists of PDB files (cf. HOBOHM et al. [9]). We also expect further improvements from replacing the polynomial fits in the potential estimation procedure by piecewise cubic fits; though at the moment it is not clear how to select the number of nodes needed to get a good but not overfitting approximation to the density. Finally, we are considering adding chirality terms to further enhance the quality of our potential. More extensive testing, e.g., using the threading test of HENDLICH et al. [8] will be done after all these enhancements have been made.

References

1. A. Bauer and A. Beyer, An improved pair potential to recognize native protein folds, Proteins: Struct. Funct. Gen. 18 (1994), 254-261.
2. A. Ben-Naim, Statistical potentials extracted from protein structures: Are these meaningful potentials? J. Chem Phys. 107 (1997), 3698-3706.
3. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E. Meyer, M.D. Bryce, J.R. Rogers, O. Kennard, T. Shikanouchi and M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures, J. Mol. Biol. 112 (1977), 535-542.
4. J.D. Bryngelson, When is a potential accurate enough for structure prediction? Theory and application to a random heteropolymer model of protein folding, J. Chem. Phys. 100 (1994), 6038–6045.
5. P.E. Gill, W. Murray and M.H. Wright, Practical optimization, Acad. Press, London 1981.
6. J. R. Gunn, A. Monge, R.A. Friesner and C.H. Marshall, Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6.2A resolution, J. Phys. Chem. 98 (1994), 702-711.
7. P. Hall and J.S. Marron, Lower bounds for bandwidth selection in density estimation, Probab. Th. Rel. Fields 90 (1991), 149-173.
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, J. Mol. Biol. 216 (1990), 167-180.
9. U. Hobohm, M. Scharf, R. Schneider and C. Sander, Selection of representative protein data sets, Protein Sci. 1 (1992), 409-417.
10. J. Holland, Genetic algorithms and the optimal allocation of trials, SIAM J. Computing 2 (1973), 88-105.
11. L. Holm and C. Sander, Database algorithm for generating protein backbone and side-chain co-ordinates from a C^α trace, J. Mol. Biol. 218 (1991), 183-194.
12. L. Holm and C. Sander, Fast and simple Monte Carlo algorithm for side chain optimization in proteins. Proteins 14 (1992), 213-223.
13. T. Huber and A.E. Torda, Protein fold recognition without Boltzmann statistics or explicit physical basis, submitted to Protein Sci. (1997).
14. M.C. Jones, J.S. Marron and S.J. Sheather, Progress in data-based bandwidth selection for kernel density estimation, Comput. Statist. 11 (1996), 337-381.
15. S. Kirkpatrick, C.D. Geddat, Jr., and M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983), 671-680.
16. J. Kostrowicki and H.A. Scheraga, Application of the diffusion equation method for global optimization to oligopeptides, J. Phys. Chem. 96 (1992), 7442-7449.
17. M. Levitt, A simplified representation of protein confomations for rapid simulation of protein folding, J. Mol. Biol. 104 (1976), 59-107.
18. M. Levitt and A. Warshel, Computer simulation of protein folding, Nature 253 (1975), 694-698.
19. O. Lund, J. Hansen, S. Brunak and J. Bohr, Relationship between protein structure and geometrical constraints, Protein Sci. 5 (1996), 2217-2225.

20. C.D. Maranas, I.P. Androulakis and C.A. Floudas, A deterministic global optimization approach for the protein folding problem, pp. 133-150 in: Global minimization of nonconvex energy functions: molecular conformation and protein folding (P. M. Pardalos et al., eds.), Amer. Math. Soc., Providence, RI, 1996.
21. J.J. Moré and Z. Wu, Global continuation for distance geometry problems, SIAM J. Optimization 7 (1997), 814-836.
22. A. Neumaier, Molecular modeling of proteins and mathematical prediction of protein structure, SIAM Rev. 39 (1997), 407-460.
23. A. Neumaier, A nonuniqueness theorem for empirical protein potentials, in preparation.
24. M. Oobatake and G.M. Crippen, Residue-residue potential function for conformational analysis of proteins, J.Phys. Chem. 85 (1981), 1187-1197.
25. T. Schlick and A. Fogelson, TNPACK – A truncated Newton minimization package for large scale problems, ACM Trans. Math. Softw. 18 (1992), 46-70; 71-111.
26. G. Schwarz, Estimating the dimension of a model, Ann. Statistics 6 (1978), 461–464.
27. D. Shortle, Y. Wang, J. Gillespie and J.O. Wrabl, Protein folding for realists: a timeless phenomenon, Prot. Sci. 5 (1996), 991-1000.
28. M.J. Sippl, Boltzmann's principle, knowledge based mean fields and protein folding, J. Comp. Aided Mol. Design 7 (1993), 473-501.
29. M.J. Sippl, M. Hendlich and P. Lackner, Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments, Protein Sci. 1 (1992), 625-640.
30. D.R. Stampf, C.E. Felser and J.L. Sussman, PDBBrowse – a graphics interface to the Brookhaven Protein Data Bank, Nature 374 (1995), 572-574.
31. S. Sun, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, Protein Sci. 2 (1993), 762-785.
32. S. Sun, Reduced representation approach to protein tertiary structure prediction: statistical potential and simulated annealing, J. Theor. Biol. 172 (1995), 13-32.
33. P.D. Thomas and K.A. Dill, Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. 257 (1996), 457-469.
34. P. Ulrich, W. Scott, W.F. van Gunsteren and A. Torda, Protein structure prediction force fields: parametrization with quasi Newtonian dynamics, Proteins 27 (1997), 367-384.
35. L.L. Walsh, Navigating the Brookhaven Protein Data Bank, Cabos Communication 10 (1994), 551-557.