# 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 $C_\alpha$ 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 Ångstrøm. 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 minima within 1.3–4.7Å of the PDB geometry, with one exception that has an error of 8.5Å.

Moreover, a nonuniqueness theorem is given that shows that no set of equilibrium geometries can determine the true effective potential energy function.

**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 [18].

[*] 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 exponential) 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.

As in the first smooth residue potential ever published (OOBATAKE & CRIPPEN [19]), we take as essential only the $C_\alpha$ atoms; this is sufficient since there are reasonably reliable methods (HOLM & SANDER [9,10]) that compute a full atom geometry from the geometry of the $C_\alpha$ atoms. 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.

Traditionally, pair potentials are determined by assuming that a set of known structures, generally taken from the Brookhaven Protein Data Bank (PDB) [2,29,24], is an equilibrium ensemble of structures, so that 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. [7]. A more detailed overview can be found in SIPPL [22]. Other empirical potential construction techniques are discussed in BAUER & BEYER [1] and ULRICH et al. [28]. 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 [25] for apamin (18 residues) and mellitin (26 residues) using genetic algorithms, by Sun [26] for mellitin, APPI (36 residues) and crambin (46 residues) using simulated annealing, by Gunn et al. [4] for myoglobin (153 residues) using a combination of simulated annealing and genetic algorithms, and by Sippl et al. [23] 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.

It is remarkable, and a fact so far seemingly not appreciated, that the approach of determining empirical potentials from equilibrium data is intrinsically limited. Indeed, we prove in Section 2 a nonuniqueness theorem that shows that the set of local and global minimizers and stationary points of any family of potentials $V(s, x)$ can be obtained from an infinite family of other potentials. Thus, no set of equilibrium geometries can determine the true effective potential energy function. While this implies that empirical potentials derived from the PDB will probably never be useful for dynamical studies, the argument also shows that if some empirical potential is able to predict correct protein folds then many other empirical potentials will do so, too. Hence the construction of empirical potentials for fold prediction is much less constrained than one might think initially.

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 $C_\alpha$-atoms at given bond distance $d$ and given residue assignments $a_1, a_2$. 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 [5] and the survey by Jones et al. [12].)

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.

## 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 $C_\alpha$ 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

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

i.e., $x_s$ is the 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., [3,20]), and can be combined with global search techniques such as simulated annealing (e.g., [13,26]), genetic algorithms (e.g., [8,25]), smoothing methods (e.g., [17,14]) or branch and bound techniques (e.g., [16]) 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. This is a consquence of the following general result that seems not to have been discussed before.

***Nonuniqueness Theorem.*** *Suppose that $V(s,x)$ is smooth, interpolates the energies*

$$V_s = V(s, x_s) \quad \textit{for all } s \in S$$

*of a set $S$ of protewin sequences, and has there* **global** *minima. Then the same holds for the modified potential*

$$\bar{V}(s,x) = V(s,x) + \frac{1}{2}\nabla V(s,x)^T G(s,x) \nabla V(s,x)$$

*for any positive semidefinite, smooth matrix valued function $G(s,x)$. Moreover, all stationary points are preserved and have the same energies.*

*Proof.* Since $G(s,x)$ is positive semidefinite, $\bar{V}(s,x) \geq V(s,x)$ for all $x$, and equality holds when $\nabla V(s,x) = 0$, i.e., at all stationary points. In particular, the global minimizers and their energies are preserved.

The gradient of the modified potential,

$$\nabla \bar{V} = \nabla V + G \nabla V + \frac{1}{2} \nabla V^T \nabla G \nabla V,$$

vanishes whenever $\nabla V = 0$; thus all stationary points are preserved in location and energy. (However, unless $G$ is small, new stationary points may be introduced, and transition points may turn into local minima.) □

Thus, no set of equilibrium geometries can determine the true effective potential energy function. This complements the findings of THOMAS & DILL [27] (through a simulation study) that statistical potentials may 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.

While this is disappointing, the argument 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. Of course, this means that the "energies" computed by an empirical potential may have little to do with real energies; but for the purpose of tertiary structure prediction, this does not matter.

## 3 The Form of the Potential

The finiteness of the database and our nonuniqueness theorem make it imperative to use qualitative theoretical assumptions in the derivation of an appropriate empirical potential function.

Our potential,

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

is a weighted sum of smooth *surface potentials* $V_a(s,x)$ that model the interaction of an amino acid with label $a$ with the solvent, and of smooth *pair potentials* $V_\gamma(s,x)$ that model interactions between pairs of amino acids classified to be of the same class $\gamma$. The weights $z_a$ and $z_\gamma$ are positive constants. (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. [15], where the surface term is a function of a smooth approximation to the number of neighbors of a $C_\alpha$ 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 $C_\alpha$ atoms, and the *pair potentials* $U_\gamma(r)$ are smooth functions of the distance with compact support $[0, r_{\max}]$. We chose the cutoff $r_{\max}$ at 12 Ångstrøm, since amino acids at most three residues apart are at distance < 12Åin known proteins.

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 $C_\alpha$ 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.

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

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

The constants were chosen such that $q_i$ has approximately mean 1 and is reasonably independent of the protein size $n$. Since $q_{ik}$ is large only when the $i$th and the $k$th residue are close, and since there are fewer residues close to a residue at the surface, $q_i$ is larger on the surface than in the interior.

$q_i$ (and hence each surface potential) is a smooth function of the coordinates, except when some position $x_k$ $(k \neq i)$ coincides with $x_i$ and $q_i$ 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

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

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 [22], 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 $\gamma$ used. (For quadratic programming in general, see, e.g., [3].)

## 4 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 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 [22].)

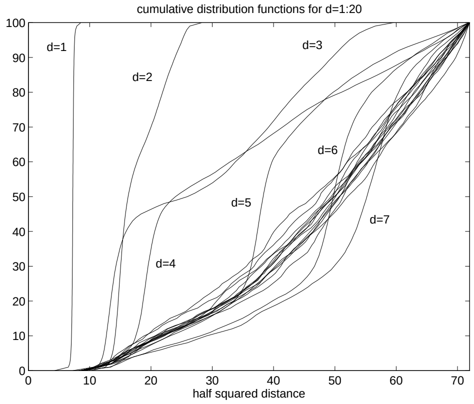

**Fig. 1.** Cumulative distribution function of half squared distances of amino acid pairs at residue distance $d$ ($d = 1, \ldots, 20$), truncated at $c_{\max} = 72$ (12 Å cutoff)

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 $O(n^{-1/2})$ 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

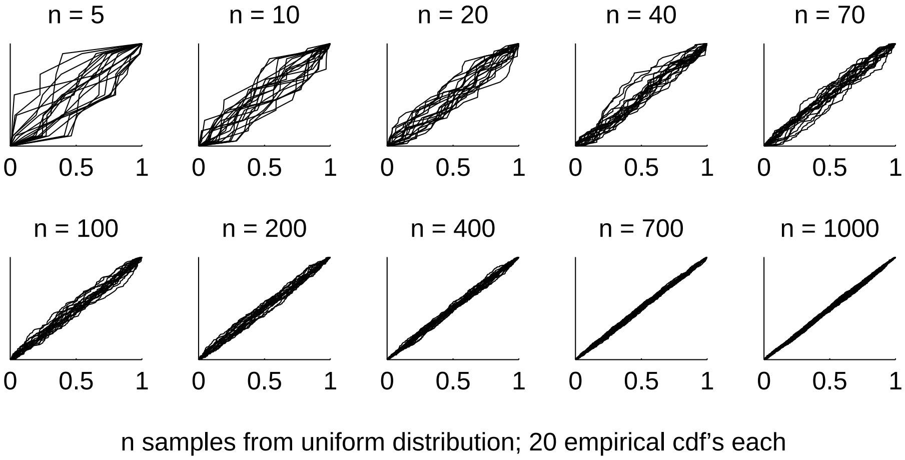

**Fig. 2.** Empirical cdf's for many samples of increasing size from a uniform distribution

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. [28] use clustering to group the 210 unordered amino acid pairs of residue distance 3 into 36 classes based on some energy measure, and Huber & Torda [11] 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 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 $n = 100$.)

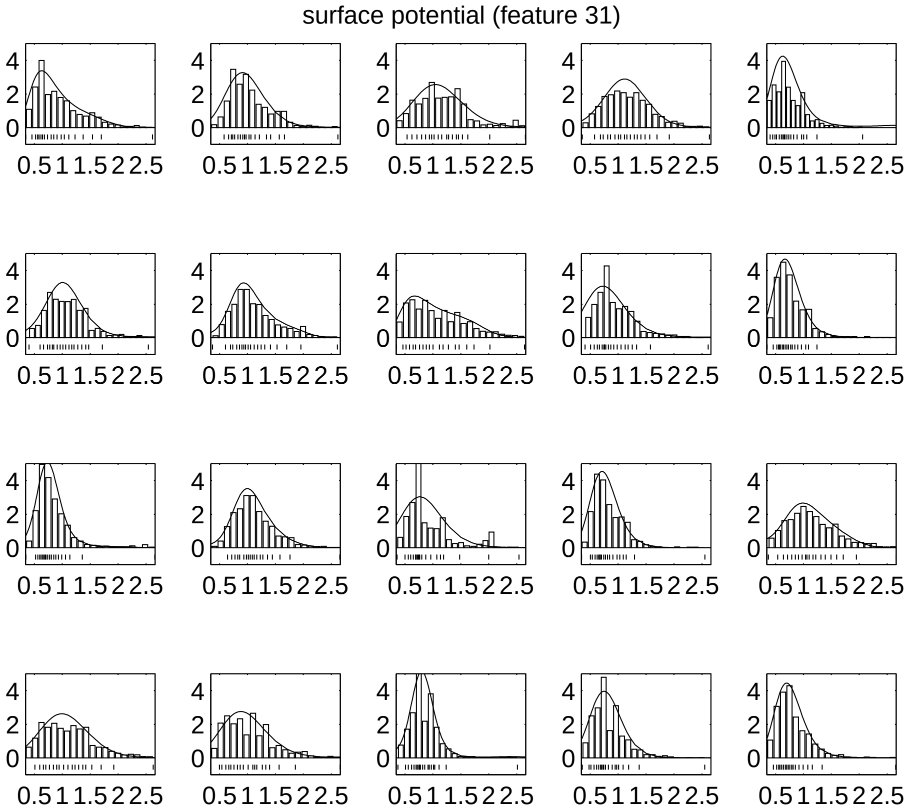

**Fig. 3.** Histograms and estimated densities $e^{-\beta W(q)}$ for surface potentials (diagrams correspond rowwise to the amino acids Ala Arg Asn Asp Cys; Gln Glu Gly His Ile; Leu Lys Met Phe Pro; Ser Thr Trp Tyr Val)

One therefore needs smooth density estimation techniques that are 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). In *http://solon.cma.univie.ac.at/~neum/stat.html#density*, some publicly available density estimators are listed, but these tend to oversmooth the densities. So we tried a number of ideas based on numerical differentiation of the empirical cdf to devise a better density estimator.

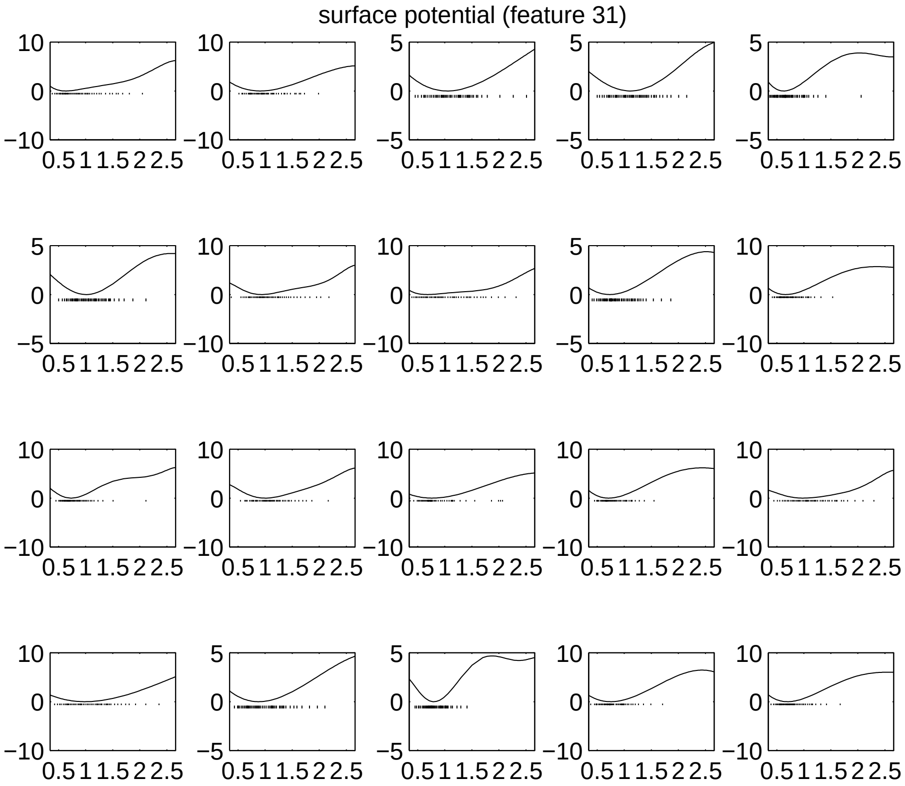

**Fig. 4.** Surface potentials $W(q)$ (diagrams correspond rowwise to the amino acids Ala Arg Asn Asp Cys; Gln Glu Gly His Ile; Leu Lys Met Phe Pro; Ser Thr Trp Tyr Val)

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 $l$th sample point in increasing order. $W(q)$ is estimated from this relation by polynomial regression, and the Bayes criterion of Schwarz [21] 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. [12] and believe that this recipe gives a reasonably reliable density estimator. It is not perfect in that it suffers occasionally from picking an unsuitable polynomial order for the potential, and it has problems when the original density is not of the form $e^{-W(q)}$ for some low degree polynomial $W(q)$.

For the surface potentials, sufficiently many data were available, and no further problems appeared. The resulting densities and potentials are shown in Figure 3 and Figure 4, respectively. The hydrophobic amino acids are easily recognized as those for which the peak is at small values of $q$.

## 5 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. [28], 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 potential has minima within 1.3–4.7Å of the PDB geometry, with one exception that has an error of 8.5Å.

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. [7]). 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

**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 |

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. [6] 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. 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.
3. P.E. Gill, W. Murray and M.H. Wright, Practical optimization, Acad. Press, London 1981.
4. 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.
5. P. Hall and J.S. Marron, Lower bounds for bandwidth selection in density estimation, Probab. Th. Rel. Fields 90 (1991), 149-173.
6. 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.
7. U. Hobohm, M. Scharf, R. Schneider and C. Sander, Selection of representative protein data sets, Protein Sci. 1 (1992), 409-417.
8. J. Holland, Genetic algorithms and the optimal allocation of trials, SIAM J. Computing 2 (1973), 88-105.
9. L. Holm and C. Sander, Database algorithm for generating protein backbone and side-chain co-ordinates from a $C^\alpha$ trace, J. Mol. Biol. 218 (1991), 183-194.

10. L. Holm and C. Sander, Fast and simple Monte Carlo algorithm for side chain optimization in proteins. Proteins 14 (1992), 213-223.
11. T. Huber and A.E. Torda, Protein fold recognition without Boltzmann statistics or explicit physical basis, submitted to Protein Sci. (1997).
12. 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.
13. S. Kirkpatrick, C.D. Geddat, Jr., and M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983), 671-680.
14. J. Kostrowicki and H.A. Scheraga, Application of the diffusion equation method for global optimization to oligopeptides, J. Phys. Chem. 96 (1992), 7442-7449.
15. O. Lund, J. Hansen, S. Brunak and J. Bohr, Relationship between protein structure and geometrical constraints, Protein Sci. 5 (1996), 2217-2225.
16. 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.
17. J.J. Moré and Z. Wu, Global continuation for distance geometry problems, SIAM J. Optimization 7 (1997), 814-836.
18. A. Neumaier, Molecular modeling of proteins and mathematical prediction of protein structure, SIAM Rev. 39 (1997), 407-460.
19. M. Oobatake and G.M. Crippen, Residue-residue potential function for conformational analysis of proteins, J.Phys. Chem. 85 (1981), 1187-1197.
20. 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.
21. G. Schwarz, Estimating the dimension of a model, Ann. Statistics 6 (1978), 461–464.
22. M.J. Sippl, Boltzmann's principle, knowledge based mean fields and protein folding, J. Comp. Aided Mol. Design 7 (1993), 473-501.
23. 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.
24. 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.
25. S. Sun, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, Protein Sci. 2 (1993), 762-785.
26. S. Sun, Reduced representation approach to protein tertiary structure prediction: statistical potential and simulated annealing, J. Theor. Biol. 172 (1995), 13-32.
27. P.D. Thomas and K.A. Dill, Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. 257 (1996), 457-469.
28. 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.
29. L.L. Walsh, Navigating the Brookhaven Protein Data Bank, Cabos Communication 10 (1994), 551-557.