Protein Science (1998), 7:142–149. Cambridge University Press. Printed in the USA.

Protein fold recognition without Boltzmann statistics or explicit physical basis

THOMAS HUBER AND ANDREW E. TORDA

Research School of Chemistry, Australian National University, Canberra, 0200 Australia

(RECEIVED May 29, 1997; ACCEPTED August 28, 1997)

Abstract

We present a fast method for finding optimal parameters for a low-resolution (threading) force field intended to distinguish correct from incorrect folds for a given protein sequence. In contrast to other methods, the parameterization uses information from $>10^7$ misfolded structures as well as a set of native sequence–structure pairs.

In addition to testing the resulting force field's performance on the protein sequence threading problem, results are shown that characterize the number of parameters necessary for effective structure recognition.

Keywords: fold recognition; low-resolution force field; optimization algorithm; parameter determination; threading

Currently, there is no shortage of low-resolution, protein fold recognition force fields (Lemer et al., 1995; Sippl, 1995; Böhm, 1996; Jernigan & Bahar, 1996; Jones & Thornton, 1996; Sippl & Flöckner, 1996; Torda, 1997). These are nearly all designed to tackle the threading problem, where a sequence is tested for compatibility with a series of structures and a pseudo-potential energy function is applied to find the most appropriate structure for some sequence.

It is not known whether a protein's fold can be explained simply by internal interactions or whether it is the result of complex interplay with the environment and folding history. Consequently, the optimal fold recognition function may not need to be based on real physical properties. Instead, it may simply reflect some common denominator among naturally expressed proteins (and solved structures).

Originally, it was seen as an achievement for a method to be able to recognize a sequence's native structure from a large number of wrong, decoy structures (Bowie et al., 1991; Jones et al., 1992). Since then, the problem of self-recognition seems to have become a minimal requirement (Defay & Cohen, 1996; Jones & Thornton, 1996). With this baseline, a new force field is probably only interesting if there is evidence of remarkable performance or some cunning innovation. The work here may not satisfy either of these criteria, but it does have some interesting properties. There is no reliance on Boltzmann statistics (Jones et al., 1992) nor on any obvious physics. Rather than merely aim for self-recognition, the methodology optimizes the statistical significance of such recognition. This is based on the philosophy of defining a criterion for force field quality and then adjusting parameters to optimize this property (Seetharamulu & Crippen, 1991; Maiorov & Crippen, 1992; Hao & Scheraga, 1996; Koretke et al., 1996; Mirny & Shakhnovich, 1996; Ulrich et al., 1997). Next, the parameterization scheme includes the effect of structures generated by threading. Unlike earlier work (Ulrich et al., 1997), a tractable scheme has been devised whereby one can easily handle parameterization with more than 300 native structures and 10^7 misfolded alternative structures. Most importantly, the force field functional forms were chosen so that one could guarantee convergence using simple gradient-based optimization. Finally, the method was applied to give some estimate of the force field's "learning capacity," or the appropriate number of adjustable parameters. This work is also based on a fundamentally unusual approach to the general problem of protein fold recognition. The work here is intended to produce a force field that only applies to native-like structures and is optimized for that purpose. This means that one accepts at the outset that it may not be the best method for sequence–structure alignment calculations and should be tested only on ungapped alignments. A separate and specifically optimized force field will be used for gapped alignments (unpubl. results).

Theory

Energy function

Each amino acid was represented by five interaction sites. Four of these were at backbone atom positions (N, C^α, C, and O) and one at the position of the C^β carbon. For glycine residues, an interaction site was positioned at the location of a fictitious C^β atom calculated assuming ideal geometry. The total energy of a chain of length N was the sum of pairwise interactions between all atoms and a solvent/environment energy of each residue based only on the C^α atom position:

$$E_{tot} = \sum_{i}^{5N} \sum_{j>i}^{5N} E_{pair}(i,j) + \sum_{k}^{N} E_{sol}(k), \quad (1)$$

Reprint requests to: Andrew E. Torda, Research School of Chemistry, Australian National University, Canberra, 0200 Australia; e-mail: andrew.torda@anu.edu.au.

where i and j are indices running over the five interaction sites in each residue and k runs over the set of C^α atoms.

The pair interaction energy $E(i,j)$ between two atoms i and j depended on spatial distance d_{ij}, topological distance s_{ij} in the amino acid sequence, and the atom types t_i and t_j. A simple sigmoidal function was chosen to describe atom pair interactions:

$$E_{pair}(i,j) = p(s_{ij},t_i,t_j)\Big(1 - \tanh(w_{pair}(d_{ij} - d_{ij}^0))\Big). \quad (2)$$

The parameters $p(s_{ij},t_i,t_j)$ determine the interaction strength and were chosen such that native sequence–structure pairs would be optimally discriminated from non-native alternative combinations. The parameters d_{ij}^0 and w_{pair} control the step position and slope of the function and were based on crude database statistics. To reduce the total number of parameters, interactions did not depend on chain direction, so $p(s_{ij},t_i,t_j) = p(s_{ij},t_j,t_i)$, and only three classes of topological distances were distinguished: $i \rightarrow j = i + 2$, $i \rightarrow j = i + 3$, and $i \rightarrow j \geq i + 4$. Each class was treated separately with different sets of parameters. Adjacent ($i \rightarrow j = i + 1$) residues were not treated explicitly in the interaction function.

The force field was based on 24 atom types. These were four backbone atoms, N, C^α, C, and O, and 20 different C^β atoms, depending on the amino acid type. This meant that, for a given topological distance, there were $\{24(24 + 1)\}/2 = 300$ interaction types.

The second term in Equation 1 was a single residue contribution depending on the C^α environment of each residue. One could see this as analogous to a solvent interaction contribution, but it was not parameterized as such and will also include general, nonspecific environment contributions. For simplicity and consistency, the same functional form as for pair interactions was used:

$$E_{sol}(i) = p(a_i)\Big(1 + \tanh(w_{sol}(n(i) - n^0))\Big), \quad (3)$$

where a_i is the amino acid type, $n(i)$ is the number amino acids separated by more than three amino acids in the sequence with a C^α carbon within 5.8 Å of C_{ij}^α. The constant n^0 was set to 3, approximately the average of $n(i)$ over all amino acid types.

This resulted in a total of 920 parameters to adjust [300 pair interaction parameters $p(s_{ij},t_i,t_j)$ per class of topological distance s_{ij} and 20 atom parameters $p(a_i)$]. In addition, a total of 47 nonadjustable parameters were used (15 d_{ij}^0 parameters per topological distance and two parameters w_{pair} and w_{sol} determining the function slopes).

Parameter optimization

Given a functional form for the discrimination function, one wants a set of parameters that optimally discriminate native from alternative sequence–structure combinations. If one assumes that the energies for one sequence on all possible nongapped alignments are Gaussian distributed, the normalization factor (z-score) describes the relative position of the native state relative to the distribution of all states:

$$z = -\frac{\langle E - E_{nat}\rangle}{\sqrt{\langle (E - E_{nat})^2\rangle - (\langle E - E_{nat}\rangle)^2}}$$

$$= -\frac{\langle \Delta E\rangle}{\sqrt{\langle \Delta E^2\rangle - (\langle \Delta E\rangle)^2}}. \quad (4)$$

E denotes the energy of an alternative structure, E_{nat} is the energy of a native structure, and the brackets $\langle\ \rangle$ denote the arithmetic average over all configurations.

In a similar vein to the idea of Koretke et al. (1996) and Hao and Scheraga (1996), the functional forms, Equations 2 and 3, were chosen such that energy was linearly dependent on $p(v)$, a scaling parameter dependent on the interaction type v. v, in turn, was given by the switching function:

$$v = \begin{cases} p(s_{ij},t_i,t_j) & \text{two-body interaction parameters (Equation 2)} \\ p(a_i) & \text{single-body interaction parameters (Equation 3).} \end{cases} \quad (5)$$

Using this definition, an energy calculation is equivalent to the evaluation of a scalar product of a parameter vector $\vec{P}$ and a vector $\vec{X}(\vec{R},\vec{S})$ depending on both coordinates $\vec{R}$ and sequence $\vec{S}$:

$$E = \vec{P}\vec{X}(\vec{R},\vec{S}). \quad (6)$$

The dimension of the vectors was given by the number of adjustable parameters n, and the vectors were defined as

$$\vec{P} = \begin{pmatrix} p(v=1) \\ p(v=2) \\ \vdots \\ p(v=n) \end{pmatrix}. \quad (7)$$

$$\vec{X}(\vec{R},\vec{S}) = \begin{pmatrix} \sum_{ij(v=1)} \Big(1 - \tanh(w_0(d_{ij} - d_0))\Big) \\ \sum_{ij(v=2)} \Big(1 - \tanh(w_0(d_{ij} - d_0))\Big) \\ \vdots \\ \sum_{ij(v=n)} \Big(1 - \tanh(w_0(d_{ij} - d_0))\Big) \end{pmatrix}. \quad (8)$$

Thus, the average alternative energy is given by

$$\langle \Delta E\rangle = \vec{P}\langle \Delta\vec{X}(\vec{R},\vec{S})\rangle$$

$$= \vec{P}\Big(\langle \vec{X}(\vec{R},\vec{S})\rangle - \vec{X}(\vec{R}_{nat},\vec{S}_{nat})\Big), \quad (9)$$

where $\vec{R}_{nat}$ and $\vec{S}_{nat}$ refer to the native coordinates and sequence.

The average energy squared can be re-written as the sum over all elements of the Hadamard product of the matrices $\underline{P}$ and $\underline{X}(\vec{R},\vec{S})$:

$$\langle \Delta E^2\rangle = \sum_{i=0}^{N}\sum_{j=0}^{N} \underline{P} \otimes \langle \Delta\underline{X}(\vec{R},\vec{S})\rangle$$

$$= \sum_{i=0}^{N}\sum_{j=0}^{N} p_i p_j \langle \Delta x_i \Delta x_j\rangle, \quad (10)$$

with $\langle \Delta x_i \Delta x_j\rangle$ the covariance elements:

$$\langle \Delta x_i \Delta x_j\rangle = \langle (x_i - x_i^{nat})(x_j - x_j^{nat})\rangle$$

$$= \langle x_i x_j\rangle - \langle x_i\rangle x_j^{nat} - x_i^{nat}\langle x_j\rangle + x_i^{nat} x_j^{nat}, \quad (11)$$

where x_i denotes the ith element in $\vec{X}(\vec{R}, \vec{S})$ and p_i is denoted the ith element in $\vec{P}$.

z-Score optimization

The goal of the parameter adjustment was to optimize the discrimination power of the force field or, in other words, to obtain the most negative z-scores (Equation 4) for proteins in the calibration set. In order to find the parameter vector that optimally discriminates n_{lib} native protein structures in the calibration set from alternative structures, the target function t was constructed as a weighted average of individual protein z-scores z_i:

$$t(\vec{P}) = \frac{1}{n_{lib}} \sum_{i}^{n_{lib}} (z_i(\vec{P}) + 15)^4. \quad (12)$$

This target function can be seen as a direct measure of force field quality or capability. The fourth power was an arbitrary choice, sharply penalizing non-native conformations. The constant of 15 could be seen as a target level and should be as large as possible. The value was determined by trial and error. The target function was minimized, with respect to force field parameters, by a quasi-Newton method (Shanno & Puha, 1976). This kind of local minimization assumes that there is only one local minimum with respect to parameters. It was not proven rigorously, but the same result was achieved from five different, random start configurations of $\vec{P}$. This target function is preferred to an arithmetic average of z-scores because proteins with z-scores closer to zero are given a higher weight and a better discrimination ability can be achieved.

A minimization algorithm such as the one used in this work may require hundreds of function evaluations of Equation 12 to converge. This means one has to recalculate the z-score for each protein in the calibration set at every step parameter optimization. The most expensive aspect of this would be those properties (averages) that depend on the large number of alternate, incorrect structures. In practice, the averages given by Equations 9 and 10 are independent of the parameters and need be calculated only once at the start of an optimization. This results in a remarkably swift method for force field parameterization.

Parameter clustering

In order to test the effect of the number of parameters on force field performance, a series of parameterization calculations were done. At each step, the parameters were grouped using a standard cluster analysis algorithm, allowing the creation of an arbitrary number of class interaction types.

Given a set of such classes, the parameterization could be repeated, resulting in a new force field with fewer parameters.

The clustering algorithm was the single linkage method (Massart & Kaufmann, 1983) and was applied within each set of parameters for a single topological distance. Thus, for example, long-range parameters were not grouped with short-range parameters.

Measure of force field performance

Two different measures of force field performance were used:

1. The z-score of each native structure in the ensemble of alternative structures.
2. The correlation between energy and distance matrix error (DME) (Havel, 1990).

$$DME = \sqrt{\frac{2}{N(N-1)} \sum_{i<j}^{N} (d_{ij} - d'_{ij})^2}, \quad (13)$$

where d_{ij} is the distance between C^α atoms i and j in one conformation and d'_{ij} is the corresponding distance in the second structure.

Results and discussion

Force field tests

Force fields were tested using two different protein sets.

1. The calibration/parameterization set, used in the optimization of parameters. This is a test of recall.
2. The test set with only low sequence homology to proteins in the calibration/parameterization set. This is a test of generalization.

The purpose of the first data set was to investigate the "learning capacity" of the force field and test whether the functional forms and the parameterization process are suitable. The second set tested the ability to generalize to unknown data and decide whether the force field reflected general protein structure information or was merely over-fitting to the data. Over-fitting would occur if the number of parameters was large relative to the information content of the training data.

Suitable alternative structures

For parameterization and analysis, we used a large number of alternative conformations to define a "reference state." These alternative structures were assumed to be "protein-like" and to reproduce an ensemble of important conformations. Using fragments from known database structures guarantees protein-like conformations in that alternative conformations will have regular secondary structure, proper backbone torsion angle distributions, and will not suffer from steric overlap. Although it is not guaranteed, the alternative structures will often be as compact as native structures. Because less compact structures were not included in the parameterization, the force field may be weak at discriminating native conformations from partly unfolded or otherwise non-native structures.

The analysis of force field capability is largely based on z-scores, which assumes a Gaussian distribution of energies of alternative, misfolded structures. Figure 1A, B, C, and D shows this distribution for the alternative structures for two example proteins. The figure also shows a semi-logarithmic plot and the fit to the expected quadratic function.

z-Score of native sequence–structure combinations

Figure 2A and B shows the force field's performance for each structure in the calibration/parameterization set and in the test set. The z-score for each native sequence–structure pair is shown as a function of protein size. Few proteins larger than 100 amino acid have a z-score higher than -5, showing high statistical confidence in the ability of the force field to pick the correct fold from all alternative structures. For smaller proteins, the z-score is closer to

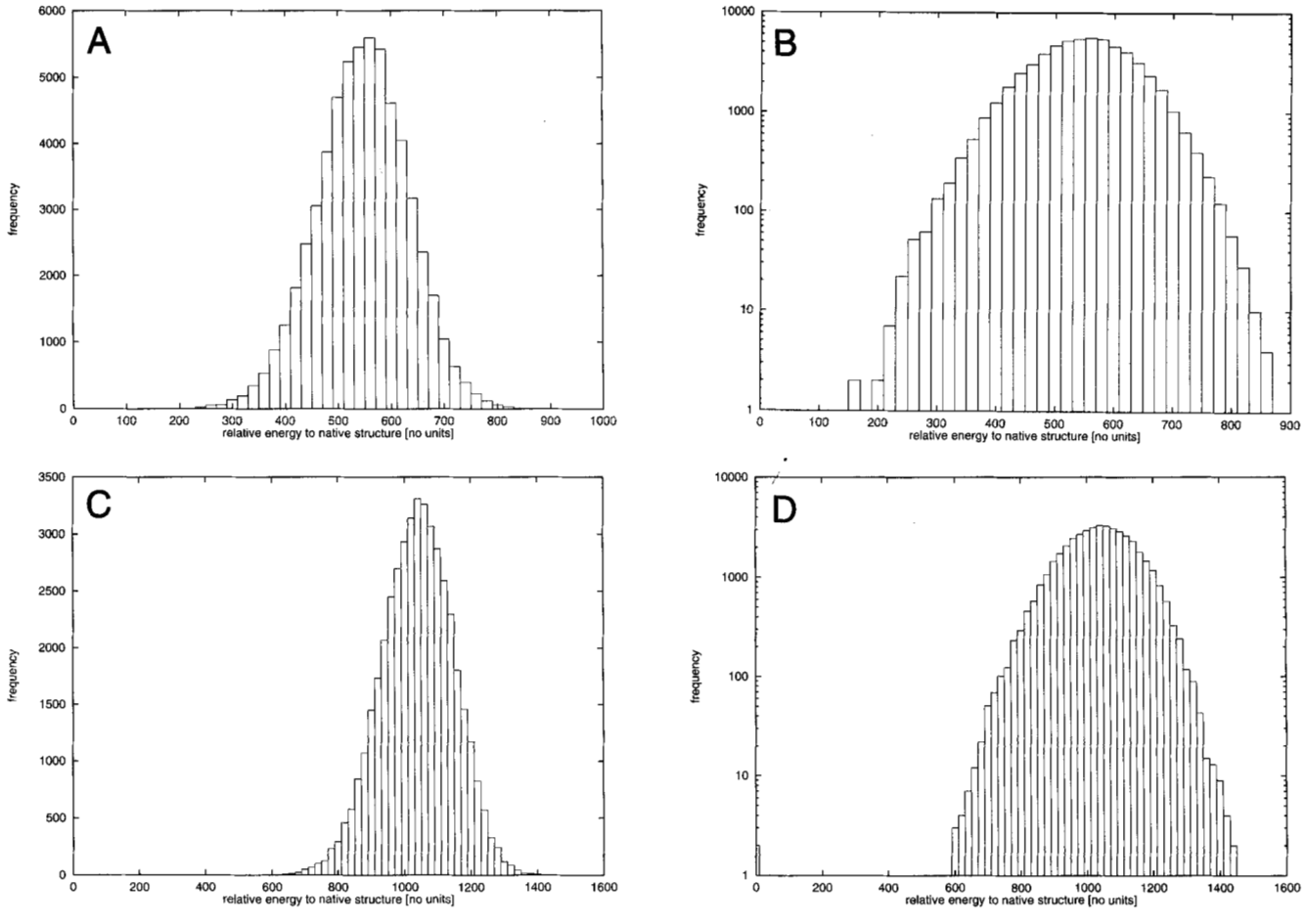

Fig. 1. Energy distribution of alternative conformations relative to energy of native structure. **A:** Protein 1etc_. **B:** Protein 1etc_, semilogarithmic plot. **C:** Protein 1aep_. **D:** Protein 1aep_, semi-logarithmic plot.

zero, meaning that one can have little confidence in predictions. This trend may be due to the larger number of (discriminative) interactions in larger molecules. It is also possible that small proteins with a relative large surface area suffer from the force field's crude accounting for solvent effects. At the same time, this trend is probably exaggerated by the smaller number of alternative structures for larger proteins. Because a sequence can only be threaded onto proteins at least as large, there are simply fewer alternatives

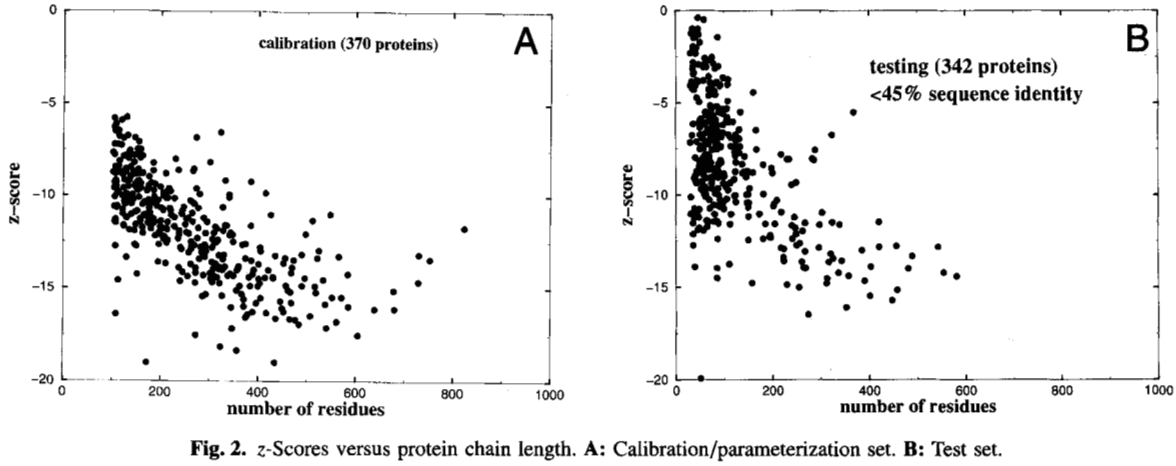

Fig. 2. z-Scores versus protein chain length. **A:** Calibration/parameterization set. **B:** Test set.

for larger structures. The poorer performance of the force field with small proteins may also be due to the nature of these proteins. Many small proteins are expressed with large pro-regions that are cleaved after folding and many contain a large number of disulfide bridges, which may force the sequence into otherwise energetically nonoptimal structures.

Despite a different protein size distribution in parameterization set and test set, the z-score results of the two different sets are very similar. This is promising for structure prediction, because the test set is unknown to the force field and demonstrates that the force field is able to generalize consistently for more than 300 proteins.

Performance with native-like structures

For successful fold recognition, it is necessary to recognize native folds and very desirable to recognize folds that are geometrically near the native. This is difficult to show in general, but an example was constructed for the trypsin fold motif. A fold library was constructed containing 103 trypsin-related proteins taken from the FSSP library (Holm & Sander, 1994) in addition to the calibration/parameterization set of proteins. This meant that a large number of similar decoy structures could be generated, as well as the alternative structures that would be available from the calibration/parameterization library. The test sequence was from β-trypsin (PDB acquisition code 1tpo), and 30,125 alternative structures were tested.

For each alternative structure, the difference from the native was measured using the DME of Equation 13 and, ideally, the force field should produce results such that alternative structures that are near to the native (low DME) are of lower energy than those very different to the native. One would expect high dispersion of energies at high DME, but at least one would want a funnel-like relationship (Bryngelson et al., 1995; Onuchic et al., 1995).

Figure 3 displays the extent to which this scoring function produces a funnel-like relation. When the structures are highly similar to trypsin (DME $\leq$ 1 Å), the energies are close to the native energy and well separated from the energy distribution of alternative structures. In this case, the fold can be recognized correctly. The limit of recognition seems to be near 4 Å when the energy of correct folds becomes too close to incorrect folds.

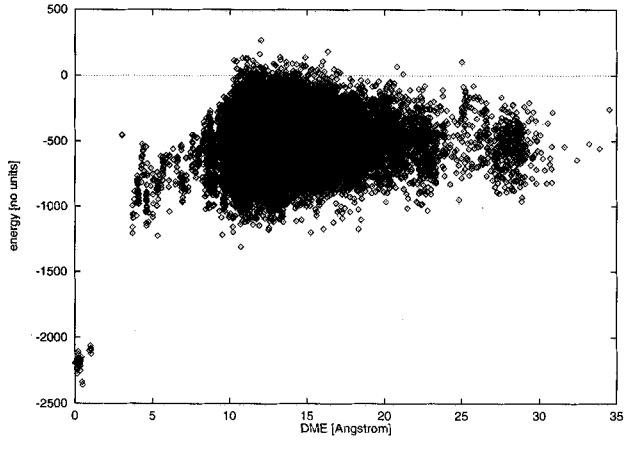

Fig. 3. Force field energies versus DME for β-trypsin. The native sequence was threaded onto 30,125 structures.

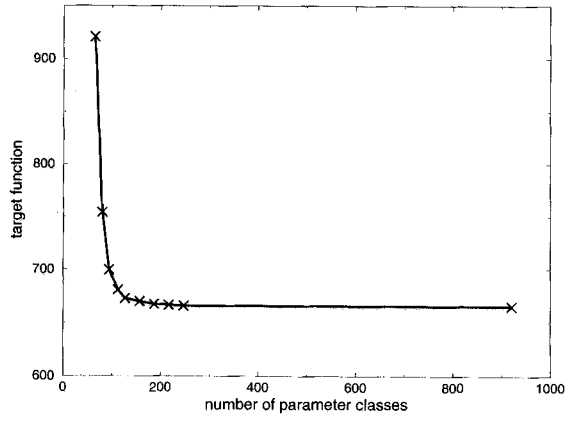

Fig. 4. Target function as a function of the number of parameter classes in the parameterization process.

Dependence on number of parameters

The number of parameters is a crucial point in any data-fitting exercise. A certain number of parameters is necessary to fit the general features of the data, but the number of parameters should be kept small to avoid over-fitting. In the case of fitting parameters for fold recognition force fields, literature estimates of the approximate number of parameters span the range from less than ten (Sun et al., 1995; Thomas & Dill, 1996) to tens of thousands (Hendlich et al., 1990; Jones & Thornton, 1993). We tried to explore the lower limit of the number of parameters in our fold recognition force field. Although the appropriate number of parameters depends on the functional forms used for interaction function, our results may be some upper limit because our interaction functions were so simple. Figure 4 summarizes the results of force field parameterization with different numbers of parameter classes. When reducing the number of classes from initial 920 to 127, the target function (Equation 12) of the parameterization hardly changes. With further clustering of classes to below $\approx$100 parameters, there is a rapid increase of the target function and significant deterioration of quality of the fit.

Ideally, the number of parameters should reflect some number of underlying degrees of freedom in the data. Unfortunately, this is an unknown quantity. It is then especially interesting to compare these results with Thomas and Dill (1996). Using different functional forms and a smaller test set, they suggested there was no benefit in using more than 10 amino acid types. Ultimately, a more direct comparison would be useful.

Is the functional form appropriate for fold recognition?

This functional form was chosen for several reasons. First, it is sufficient to distinguish two states and similar "contact potential terms" have a successful history in protein fold prediction (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985) and lattice simulations (Lau & Dill, 1989; Shakhnovich & Gutin, 1989, 1990). Second, the use of a continuous form with a defined derivative with respect to parameters allows the use of efficient minimization methods. Last, the form does not have a narrow range of distances that result in low energy, as would be the case with a Lennard–Jones like term (Levitt, 1976; Oobatake & Crippen, 1981). This means that calculated pseudo-energies are less sensitive to pertur-

Table 1. *Parameterization protein set (PDB acquisition codes)*

121p_,	131l_,	153l_,	193l_,	1aaj_,	1aak_,	1abrB,	1add_,	1adeA,	1aep_,	1aliA,	1amg_,
1amp_,	1aorA,	1aozA,	1arb_,	1ars_,	1ash_,	1asu_,	1atlA,	1atpE,	1ayaA,	1bbpA,	
1bbt2,	1bbt3,	1bcfA,	1bdmB,	1bec_,	1bet_,	1bglA,	1bip_,	1bnh_,	1bp2_,	1briC,	
1bucA,	1bvp1,	1bw4_,	1byb_,	1cauB,	1cbg_,	1ccr_,	1celA,	1cfb_,	1chd_,	1chmA,	1cid_,
1clc_,	1cmbA,	1cnsA,	1cpcA,	1cpcB,	1crl_,	1csh_,	1csn_,	1ctn_,	1ctt_,	1cus_,	1cyg_,
1daaA,	1deaA,	1dhr_,	1dih_,	1dlc_,	1dlhA,	1dlhB,	1dpb_,	1dppA,	1dsbA,	1dupA,	
1dynA,	1dyr_,	1eca_,	1ede_,	1eft_,	1eriA,	1esc_,	1etc_,	1exh_,	1fbaA,	1fc2D,	1fcdA,
1fcdC,	1fkj_,	1fnc_,	1fnf_,	1fps_,	1fruA,	1gcb_,	1ghr_,	1gky_,	1gln_,	1gmfA,	1gof_,
1gpb_,	1gpc_,	1gph1,	1gpr_,	1gseA,	1han_,	1hbq_,	1hcd_,	1hdcA,	1hdg0,	1hfh_,	
1hgeA,	1hjrA,	1hlb_,	1hleA,	1hmpA,	1hmt_,	1hmy_,	1hngA,	1hpm_,	1hslA,	1htbA,	
1htmD,	1htp_,	1hucB,	1hvd_,	1hvm_,	1hxn_,	1i1b_,	1iae_,	1ikfL,	1ilk_,	1irk_,	1irl_,
1iscA,	1ivd_,	1knb_,	1kptA,	1krt_,	1lba_,	1lcpA,	1lct_,	1ldm_,	1lfaA,	1lgaA,	1lis_,
1lki_,	1lpbB,	1lpe_,	1ltsA,	1ltsD,	1lxa_,	1mat_,	1mhcA,	1mhlA,	1mhlC,	1minB,	
1mioA,	1mldA,	1mls_,	1mml_,	1mmoB,	1mmoD,	1mmoG,	1mnc_,	1mpp_,	1mrj_,		
1msaA,	1msc_,	1msfC,	1mup_,	1nal1,	1nar_,	1nbaA,	1ncfA,	1ndh_,	1nfp_,	1nhkL,	
1nif_,	1nipA,	1omp_,	1onc_,	1opr_,	1ora_,	1ordA,	1osa_,	1oxa_,	1oyc_,	1pbe_,	1pbn_,
1pbp_,	1pbxA,	1pdnC,	1pfkA,	1pgs_,	1phg_,	1phr_,	1pii_,	1pil_,	1pkm_,	1pkp_,	1plq_,
1pls_,	1pne_,	1pnrA,	1poc_,	1poxA,	1ppi_,	1ppn_,	1prcC,	1prcL,	1prcM,	1prr_,	1prtB,
1prtD,	1psdA,	1pspA,	1put_,	1pvc2,	1pyaB,	1pyp_,	1qorA,	1rcb_,	1rcf_,	1rci_,	1regX,
1rfbA,	1ribA,	1rpa_,	1rsy_,	1rtm1,	1rtp1,	1rvaA,	1sacA,	1sat_,	1sbp_,	1scs_,	1scuA,
1scuB,	1sesA,	1sltB,	1smnA,	1snc_,	1std_,	1sxcA,	1tahA,	1tca_,	1thv_,	1thx_,	1tlk_,
1tml_,	1tnrA,	1tph1,	1trb_,	1trkA,	1trrA,	1tssA,	1ttbA,	1tupB,	1tys_,	1udg_,	1vcaA,
1vhh_,	1vil_,	1vmoA,	1vsgA,	1was_,	1whtA,	1whtB,	1wsyB,	1xnb_,	1xylA,	1xyzA,	
1yptB,	1ytbA,	256bA,	2abk_,	2acg_,	2acq_,	2ak3B,	2alp_,	2ayh_,	2azaA,	2bbkH,	
2bltA,	2bpa1,	2bpa2,	2btfA,	2cas_,	2cba_,	2ccyA,	2cdv_,	2chr_,	2chsA,	2cpl_,	2ctc_,
2cwgA,	2cyp_,	2dkb_,	2dln_,	2dri_,	2ebn_,	2end_,	2er7E,	2fal_,	2fd2_,	2gbp_,	2gdm_,
2gstA,	2hbg_,	2hhmA,	2hmx_,	2hmzA,	2hnq_,	2hpdA,	2kauB,	2kauC,	2liv_,	2madL,	
2mev1,	2mnr_,	2mtaC,	2nacA,	2olbA,	2omf_,	2pf1_,	2pgd_,	2phy_,	2pia_,	2pnb_,	
2polA,	2por_,	2prd_,	2rn2_,	2rslB,	2sas_,	2scpA,	2sil_,	2snv_,	2stv_,	2tgi_,	2tmdA,
2tmvP,	3aahA,	3cd4_,	3chy_,	3cla_,	3est_,	3gapB,	3gly_,	3grs_,	3pgk_,	3pgm_,	3pmgA,
3pte_,	3rubS,	3sdhA,	3sicI,	3tgl_,	4blmA,	4enl_,	4fgf_,	4fxn_,	4gcr_,	4rhv1,	4rhv3,
4sbvA,	6taa_,	7icd_,	7rsa_,	8abp_,	8acn_,	8atcA,	8atcB,	8catA,	8tlnE,	9rnt_	

Table 2. *Test protein set (PDB acquisition codes)*

1aapA,	1ab2_,	1aba_,	1aboA,	1abrA,	1acp_,	1adn_,	1adr_,	1afp_,	1agt_,	1akp_,	1aky_,
1aml_,	1amy_,	1ang_,	1apa_,	1arv_,	1aszA,	1ate_,	1atx_,	1avdA,	1babB,	1bba_,	1bbhA,
1bbl_,	1bbrH,	1bfmA,	1bgh_,	1bmtA,	1bnb_,	1bovA,	1btl_,	1c5a_,	1cbn_,	1cbs_,	1cc5_,
1ccf_,	1cd8_,	1cdb_,	1cfh_,	1cgmE,	1cgo_,	1chc_,	1chl_,	1cis_,	1ckaA,	1clh_,	1coe_,
1colA,	1cot_,	1cpy_,	1crb_,	1cseE,	1cskA,	1ctaA,	1ctf_,	1ctl_,	1ctm_,	1cxc_,	1cyj_,
1d66A,	1dctA,	1dec_,	1dfnA,	1dmc_,	1eciB,	1ecmA,	1edt_,	1ego_,	1ehs_,	1eny_,	1erd_,
1erl_,	1erp_,	1esl_,	1ezm_,	1f3g_,	1fas_,	1fc2C,	1fca_,	1fct_,	1fipA,	1fivA,	1flp_, 1fosE,
1fosF,	1ftpA,	1ftt_,	1ftz_,	1fxd_,	1fxrA,	1gal_,	1gauA,	1gbrA,	1gdhA,	1gfd_,	1ghc_,
1glqA,	1gluA,	1gmpA,	1gps_,	1gsq_,	1gta_,	1hcgB,	1hcnB,	1hcrA,	1hip_,	1hks_,	1hlcA,
1hlpA,	1hma_,	1hml_,	1hnr_,	1hoe_,	1hph_,	1hpi_,	1hpt_,	1hrf_,	1hsbB,	1hsn_,	1hstA,
1htrP,	1hucA,	1hueA,	1hulA,	1humA,	1hylA,	1hyp_,	1ica_,	1iceA,	1idm_,	1idsA,	
1ifc_,	1ifj_,	1igd_,	1ikm_,	1ilr1,	1isuA,	1ithA,	1kdu_,	1knt_,	1lac_,	1ldnA,	1leb_, 1led_,
1lenA,	1lenB,	1lldA,	1lmb3,	1lmwB,	1lpbA,	1ltsC,	1lvl_,	1lybA,	1lybB,	1lyp_,	1mal_,
1mdaH,	1mdc_,	1mdkA,	1mdyA,	1mioB,	1mjc_,	1mngA,	1mnp_,	1mntA,	1mylB,		
1ner_,	1nhp_,	1npk_,	1nscA,	1ntn_,	1ntr_,	1oaw_,	1paa_,	1paz_,	1pbxB,	1pce_,	1pch_,
1pcrH,	1pdc_,	1pfiA,	1pht_,	1pi2_,	1pk4_,	1plfB,	1pmc_,	1pmlA,	1pmy_,	1pnh_,	1pod_,
1pp2L,	1ppbL,	1ppfE,	1ppt_,	1prcM,	1prhA,	1prn_,	1psf_,	1psm_,	1ptf_,	1ptq_,	1pvc3,
1pyaA,	1pyiA,	1r1a1,	1r69_,	1rblM,	1ret_,	1rip_,	1ris_,	1rpo_,	1rtc_,	1safA,	1sap_,
1scmA,	1scmB,	1scmC,	1sgt_,	1shaA,	1shfA,	1shg_,	1spbP,	1spf_,	1stu_,	1sxl_,	1tap_,
1ten_,	1tfi_,	1tfs_,	1thg_,	1tib_,	1tif_,	1tig_,	1tin_,	1tlfA,	1tnfA,	1tnn_,	1tnt_, 1treA,
1try_,	1tvt_,	1ubi_,	1ukz_,	1utg_,	1wapA,	1wfbA,	1yrnA,	1yrnB,	1ysaC,	1zaq_,	2achB,
2apr_,	2bb2_,	2bds_,	2bpa3,	2cbh_,	2crd_,	2cy3_,	2cyr_,	2exo_,	2fcr_,	2fx2_,	2fxb_,
2hpqP,	2hsp_,	2ifo_,	2kauA,	21hb_,	2mcm_,	2mev2,	2mev3,	2mev4,	2pcdA,	2pec_,	
2pleA,	2ptl_,	2sga_,	2sh1_,	2sn3_,	2spcA,	2tbvA,	2tprA,	2trxA,	2uce_,	3aahB,	3adk_,
3blm_,	3c2c_,	3egf_,	3mddA,	3rp2A,	3sgbI,	451c_,	4dfrA,	4icb_,	4ptp_,	4sgbI,	4tgf_,
5znf_,	6insE,	7aatA,	7apiA,	7apiB,	7pti_,	8dfr_,	8fabA,	8rxnA			

bations of coordinates and more tolerant of "errors" in geometry, which makes processing of structures to remove strains (Park & Levitt, 1996; Ulrich et al., 1997) unnecessary.

Is the parameterization methodology appropriate?

Parameters were optimized using function minimization, which is known to fail if the function hyper-surface is rough and barriers separate local minima. To our surprise, this is not the case with our target function (Equation 12). The hyper-surface turned out to be smooth and function minimization was able to locate a single minimum from five randomly chosen starting configurations of $\vec{P}$. As a last check, parameter dynamics simulation on the target function hyper-surface, similar to the method of Ulrich et al. (1997), combined with simulated temperature annealing (Kirkpatrick et al., 1983), was performed. No improvement of the results was obtained.

Are the measures of significance appropriate?

We used two different measures to assess the quality of the force field. Although these measurements are not completely independent of each other, they give a representative description of the performance of a force field.

The z-score describes the energy of the native or any other single structure relative to an idealized energy distribution of all alternative structures. Therefore, the z-score depends highly on quality of the decoy structures used in the ensemble of alternates.

Analyzing the correlation between energy and DME of misfolded structures avoids this problem. Nevertheless, this measure is not invariant with arbitrary scaling of energies (which does not affect discrimination properties) and therefore requires an additional external energy reference.

Concluding remarks

The philosophy of this work has been that, by building a force field optimized for discrimination, one should obtain better performance than using a force field that simply reflects structural properties of native structures. Because we find a global minimum to the force field construction problem, one could go so far as to say that this is the best possible force field. Of course, this is only true in the framework of the functional form and with respect to the target function (Equation 12) used.

The parameterization and testing has only been conducted on ungapped alignments and it is likely that the parameter set here is not ideal for calculating sequence–structure alignments. Continuing in this vein of specialized force fields, a different optimization methodology has been used with different functional forms and a different target function for the sequence–structure alignment problem (unpubl. results). This scoring function is only expected to be used, given a reasonable native-like alignment produced by a separate alignment procedure. Hence, this work has been restricted to ungapped alignments (with a very large number of decoys) and testing with common literature measures such as z-scores.

For comparison, the results here have been based on common literature measures such as z-score, and the results cover a large range of proteins. At the same time, the family of native proteins and decoys is probably different from every other in the literature, so a true comparison of force field performance is not really possible and some of the claims in this work remain as unsubstantiated as others in the literature.

Methods

Data sets

The calibration/parameterization set of proteins was taken from Hobohm and Sander (1994), March 1996 release, and consisted of protein chains such that no sequence had more than 25% sequence identity with any other member of the set. From this list, chains of 100 or more residues and with all backbone heavy atoms were selected. This resulted in a set of 370 protein chains in the final calibration set (Table 1).

A library of misfolded, decoy structures for each protein chain was generated by threading the native sequence onto all structures of the same or larger size in the protein library, resulting in a total of 10.54 million alternative structures for the 370 native structures.

For testing, a second set of 342 proteins was chosen (Table 2), again from the list of Hobohm and Sander (1994), such that no protein had more than 45% sequence identity with any other member or any member of the calibration set. Unlike the calibration set, no criterion was applied for chain length and the test set contained many chains of less than 100 residues. Using threading, there were $5.41 \cdot 10^6$ alternative, misfolded decoy structures generated for this set.

Building the force field

Parameters were optimized by minimization of the target function (Equation 12) using the parameterization set of 370 proteins until the energy gradient was smaller than 10^{-6}. After the raw force field with 920 adjustable parameters was obtained, parameters were clustered and new force fields were generated with 65, 80, 95, 112, 127, 157, 187, 217, and 247 distinct parameter classes.

Supplementary material in Electronic Appendix

The force field parameters are provided as supplementary material. A program capable of scoring alignments using this force field can be found at ftp://ftp.rsc.anu.edu.au/~torda/README.

References

Böhm G. 1996. New approaches in molecular structure prediction. *Biophys Chem 59*:1–32.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct Funct Genet 21*:167–195.

Defay TR, Cohen FE. 1996. Multiple sequence information for threading algorithms. *J Mol Biol 262*:314–323.

Hao MH, Scheraga HA. 1996. How optimization of potential functions affects protein folding. *Proc Natl Acad Sci USA 93*:4984–4989.

Havel TF. 1990. The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: A study of 1830 independently computed conformations. *Biopolymers 29*:1565–1585.

Hendlich M, Lackner P, Weitckus S, Flöckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. 1990. Identification of native protein folds among a large number of incorrect models: The calculation of low energy conformations from potentials of mean force. *J Mol Biol 216*:167–180.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Holm L, Sander C. 1994. The fssp database of structurally aligned protein fold families. *Nucleic Acids Res 24*:206–210.

Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Jones DT, Thornton JM. 1993. Protein fold recognition. *J Comp Aided Mol Design 7*:439–456.

Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr Opin Struct Biol 6*:210–216.

Kirkpatrick S, Gelatt CD Jr, Vecchi MP. 1983. Optimization by simulated annealing. *Science 220*:671–680.

Koretke KK, Luthey-Schulten Z, Wolynes PG. 1996. Self consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci 5*:1043–1059.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*:3986–3997.

Lemer CMR, Rooman MJ, Wodak SJ. 1995. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins Struct Funct Genet 23*:337–355.

Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol 104*:59–107.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct fold of globular proteins. *J Mol Biol 227*:876–888.

Massart DL, Kaufmann L. 1983. *The interpretation of analytical chemical data by the use of cluster analysis*. New York: John Wiley & Sons.

Mirny LA, Shakhnovich EI. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol 264*:1164–1179.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi chemical approximation. *Macromolecules 18*:534–552.

Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. 1995. Toward an outline of the topography of a realistic protein folding funnel. *Proc Natl Acad Sci USA 92*:3626–3630.

Oobatake M, Crippen GM. 1981. Residue–residue potential function for conformational analysis of proteins. *J Phys Chem 85*:1187–1197.

Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J Mol Biol 258*:367–392.

Seetharamulu P, Crippen GM. 1991. A potential function for protein folding. *J Math Chem 6*:91–110.

Shakhnovich EI, Gutin AM. 1989. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem 34*:187–199.

Shakhnovich EI, Gutin AM. 1990. Enumeration of all compact conformations of copolymers with random sequence of links. *J Chem Phys 93*:5967–5971.

Shanno DF, Puha KH. 1976. Minimization of unconstrained multivariate functions. *AC MTOMS 2*:87–94.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Sippl MJ, Flöckner H. 1996. Threading thrills and threats. *Structure 4*:15–19.

Sun S, Thomas PD, Dill KA. 1995. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng 8*:769–778.

Tanaka S, Scheraga HA. 1976. Medium- and long-range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules 9*:945–950.

Thomas PD, Dill KA. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA 93*:11628–11633.

Torda AE. 1997. Perspectives on protein fold recognition. *Curr Opin Struct Biol 7*:200–205.

Ulrich P, Scott WRP, van Gunsteren WF, Torda AE. 1997. Protein structure prediction force fields: Parametrization with quasi newtonian dynamics. *Proteins Struct Funct Genet 27*:367–384.