Protein Structure Prediction Force Fields: Parametrization With Quasi-Newtonian Dynamics

Patrick Ulrich,[1] Walter Scott,[1] Wilfred F. van Gunsteren,[1] and Andrew E. Torda[2]*
[1]*Computational Chemistry (Physical Chemistry) ETH Zentrum, 8092 Zürich, Switzerland*
[2]*The Research School of Chemistry, The Australian National University, Canberra ACT 0200, Australia*

ABSTRACT **We present an unusual method for parametrizing low-resolution force fields of the type used for protein structure prediction. Force field parameters were determined by assigning each a fictitious mass and using a quasi-molecular dynamics algorithm in parameter space. The quasi-energy term favored folded native structures and specifically penalized folded nonnative structures. The force field was generated after optimizing less than 70 adjustable parameters, but shows a strong ability to discriminate between native structures and compact misfolded alternatives. The functional form of the force field was chosen as in molecular mechanics and is not table-driven. It is continuous with continuous derivatives and is thus suitable for use with algorithms such as energy minimization or newtonian dynamics. Proteins 27:367–384, 1997.**
© 1997 Wiley-Liss, Inc.

Key words: protein folding; force field; structure prediction; molecular dynamics

INTRODUCTION

Ideally, one would like to be able to predict protein structures based only on the amino acid sequence and without recourse to experiment. This may not be realistic, but it may well be feasible to recognize when a new sequence will adopt a known fold[1–4] or whether a proposed structure is totally misfolded.[5] To do this, one needs some function of coordinates that yields a number reflecting the fit of sequence and structure. In molecular mechanics nomenclature, this is a potential energy, but it might also be referred to as a score[6] or profile.[7] The incentive to develop such useful potential energy functions continually increases as more protein sequences are determined and as the first complete genomes are sequenced.

To be practical, such potential energy functions are based on simple representations of proteins using only one or two interaction sites per residue. Beyond that, there is a multitude of approaches. For example, functions may be defined in discrete (usually lattice) coordinates[8–12] for the sake of speed or in continuous space for detail. In this work, we concentrate only on continuous force fields. A less obvious, but conceptually important distinction can be made on the philosophy of the force field. Some force fields can be seen as tightly connected to an underlying physical basis in that they should in some way be an average over physical terms.[13–17] Alternatively, a force field may be statistical in nature. Although it will be some kind of average of the underlying physical interactions, the functional forms may bear little relation to common terms such as electrostatic or Lennard-Jones potential energy functions. Because the functional forms are not physically based, they may only be useful for a narrow range of structures.

The force field in this work is of the simple statistical type, which immediately limits its ability to generalize. The parametrization procedure is based on native and other compact conformations so one would not expect it to extrapolate to unfolded conformations. One can also see that, although we use nomenclature from molecular mechanics, our energies bear no relation to any real potential energy.

Given these caveats, the aims of this force field parametrization can be stated. The calculations are based on a wide variety of proteins and should encapsulate properties common to proteins in general. The parameters will only be relevant to compact structures, but they should be useful across many sequences and folded conformations. The ability to generalize is further enhanced by using a small number (<70) of adjustable parameters so as to avoid fitting to noise in the dataset.[18] This relatively small number of parameters was brought about by using only four kinds of interaction and grouping interaction types (not amino acids) into classes.

The method of parametrization in this work is quite unusual in the area of low-resolution force fields. The procedure began by taking the idea of Crippen and coworkers[19,20] that the force field should specifically penalize misfolded structures. This idea was then cast into the form of a continuous pseudoenergy function. The parameters were then given

*Correspondence to: Dr. Andrew E. Torda, The Research School of Chemistry, The Australian National University, Canberra, 0200, Australia. E-mail: andrew.torda@anu.edu.au.

Received 8 March 1996; accepted 13 August 1996.

TABLE I. Relative Weights of Force Field Terms

Term[a]	Value[b]	Interaction type
$K^{n,n+2}$	2.0	Pseudoangle
$K^{n,n+3}$	1.0	Pseudotorsion
K^{long}	0.1	Long range

[a] Refers to Equation (1).
[b] Dimensionless units.

fictitious masses, treated like particles, and subjected to quasi-newtonian dynamics. This dynamics simulation was solely used as an optimization process to improve initial estimates of parameters and was in no way a physical simulation.

METHODS

The following sections describe the protein model, interaction functions and parameters, and a methodology for optimizing the parameters. To make this interpretable, certain conventions are used. Superscripts are used to specify topological proximity, so $E^{n,n+1}$ may refer to an interaction energy of two adjacent amino acids and $\sigma^{n,n+3}$ would refer to a σ parameter for two residues with two intervening residues. Lowercase subscripts refer to any pair of residues, so r_{ab} is the distance between residues a and b, regardless of their type. Force field parameters depend on topological proximity and amino acid type, so one also needs a convention for specifying the type or class of amino acid. An uppercase subscript is used here, so ϵ_J is a parameter for residues of type J, and $\sigma_{XY}^{\mathrm{long}}$ is the σ parameter for interactions between residues of types X and Y at a long topological distance.

Protein Model and Interaction Function

Energies were calculated as if they were true potential energies, so the nomenclature is taken from conventional molecular mechanics.

Each amino acid was represented by a single point located at the Cα atom. These points then interacted in one of four ways depending on their topological proximity. The total potential energy of a protein was calculated by summing four terms:

$$E_{\mathrm{total}} = K^{n,n+1} \sum_{\substack{\text{pseudo}\\ \text{bond}}} E^{n,n+1} + K^{\mathrm{n},n+2} \sum_{\substack{\text{pseudo}\\ \text{angle}}} E^{n,n+2} + K^{n,n+3} \sum_{\substack{\text{pseudo}\\ \text{torsion}\\ \text{angles}}} E^{n,n+3} + K^{\mathrm{long}} \sum_{\substack{\text{other}\\ \text{pairs}}} E^{\mathrm{long}} \qquad (1)$$

Each of the four terms was weighted by an arbitrary constant, K, as given in Table I.

The interaction between adjacent residues is analogous to a bond potential energy in a normal force field, although there is no real bond there. This gives rise to the first term, $E^{n,n+1}$, in Equation (1). Similarly, one can define an energy term that depends on the angle included by three adjacent residues. This is the second term in Equation (1). Residues separated by two intervening connected residues form a pseudotorsion angle and their interaction gives rise to the $E^{n,n+3}$ term in Equation (1). All other amino acid pairs interact by a fourth kind of interaction, E^{long}. Each of the terms in Equation (1) had a different dependence on either parameters or amino acid types, and each is described below.

Although some initial tests were performed using a simple harmonic form for the pseudobond term, this was replaced in final implementations. We used the iterative algorithm, SHAKE[21] to hold pseudobonds to 3.81 Å, the typical distance between C$^{\alpha}$ atoms in adjacent residues with *trans* peptide bonds.[12,16] This concedes some error due to the small fraction of *cis* peptide bonds. The effect of this was that energy from a pseudobond term was redistributed into the other terms of the force field.

A Lennard-Jones-like term was used for residues separated by one intervening residue in the sequence. Considering three sequential residues i, j, and k of types I, J, and K, the interaction energy $E_{ik}^{n,n+2}$ depended on the type of central residue J and the distance r_{ik} between residues i and k. The interaction energy was calculated according to

$$E_{ik}^{n,n+2}(r_{ik}, \epsilon_J^{n,n+2}, \sigma_J^{n,n+2}) = \epsilon_J^{n,n+2} \left[5 \left(\frac{\sigma_J^{n,n+2}}{r_{ik}} \right)^{12} - 6 \left(\frac{\sigma_j^{n,n,+2}}{r_{ik}} \right)^{10} \right]. \qquad (2)$$

The energy has been written explicitly as a function of force-field parameters, as well as the distance, to highlight the fact that these were later treated as adjustable quantities. This has been deliberately cast so as to highlight the similarity with a Lennard-Jones term in a molecular mechanics force field.[22] $\sigma_J^{n,n+2}$ gives the distance of lowest energy and $\epsilon_J^{n,n+2}$ gives the depth of the energy well at that distance. The single uppercase subscript J reflects the fact that the parameters for the n, $n + 2$ interaction depend only on the type of the central residue J. When calculating the energy of a protein, the contribution from Equation (2) was summed over every sequential triplet of residues not spanning a chain break.

The same functional form was used for residues separated by two intervening residues, but with a different set of parameters and different dependence on parameters. Given four sequential residues i, j, k, and l of types I, J, K, and L, the energy $E_{il}^{n,n+3}$ depended on the types of the central pair of residues

J and K, and the distance r_{il} between the outer residues. The interaction energy was calculated according to

$$E_{il}^{n,n+3}(r_{il}, \epsilon_{JK}^{n,n+3}, \sigma_{JK}^{n,n+3}) = \epsilon_{JK}^{n,n+3}\left[5\left(\frac{\sigma_{JK}^{n,n+3}}{r_{il}}\right)^{12} - 6\left(\frac{\sigma_{JK}^{n,n,+3}}{r_{il}}\right)^{10}\right]. \quad (3)$$

This term's contribution to the total energy of the protein was calculated by summing over all sets of four sequential residues that did not span a chain break.

The last component of the potential energy, E^{long} accounted, in a mean manner, for the tendency of certain kinds of residues to adopt certain distances

$$E_{ij}^{\text{long}}(r_{ij}, \epsilon_{IJ}^{\text{long}}, \sigma_{IJ}^{\text{long}}) = \frac{1}{11}\epsilon_{IJ}^{\text{long}}\left[\left(\frac{\sigma_{IJ}^{\text{long}}}{r_{ij}}\right)^{12} - 12\left(\frac{\sigma_{IJ}^{\text{long}}}{r_{ij}}\right)^{10}\right]. \quad (4)$$

This term was summed over every pair of residues separated by more than three pseudobonds, but within a cutoff of 15 Å. This included residues within a protein, across subunits of polymeric proteins, and any bound small peptides. The choice of cutoff size was arbitrary, but ensured that, beyond a certain size, the total energy would have a linear dependence on the size of a protein.

No term was used for disulfide bridges, since it was intended to use the force field when the location of disulfide bridges would not be known. Their structural effect was only included in a mean manner along with all the other influences on parameters for Equation (1).

Protein Set For Parametrization

For force field development, one requires a set of calibration proteins. This set should be large, but contain only reliable, well-defined structures without sequence or structural homology between members.

The basis for the calibration set was provided by Hobohm and Sander[23] and consisted of proteins with less than 35% sequence homology. Structures with crystallographic R factors greater than 0.25 or resolutions worse than 2.8 Å were removed. Proteins with missing internal C^{α} coordinates were deleted, but those missing up to 7 C- or N-terminal residues were included. Decisions about the resolution of NMR structures were avoided by omitting them entirely. Similarly, any proteins with large prosthetic groups or internal metal ions were omitted. This provided an initial set of 108 proteins.

A further selection step was applied by calculating the energy of each protein using the initial crude parameter set described below. This suggested that four proteins (1cid, 1omf, 1fai, and 1tnf) were of high energy (data not shown). The energy was then decomposed into the individual terms of Equation (1). In each case, the high energy could be attributed to an unusually short (≤4.51 Å) n, $n + 2$ Cα—Cα distance. These entries were removed from the reference protein set, even though they met the criteria of resolution and R factor. Figure 1 shows the potential energy for each protein in the set as a function of size. Energies were calculated using the unrefined force field and serve to show that there were no obvious outliers or incorrect structures in the calibration set.

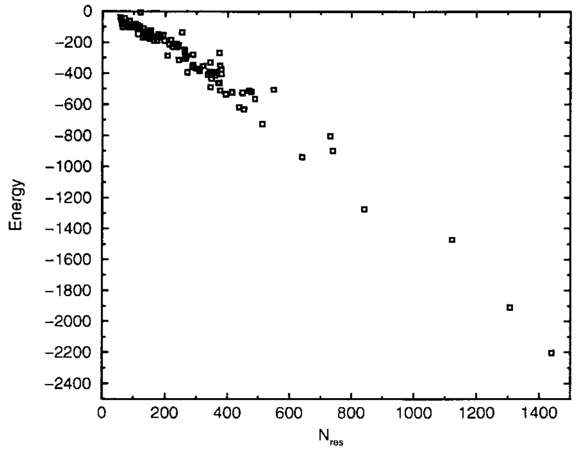


Fig. 1. Total energy as a function of protein size. Energy is in arbitrary units. N_{res} refers to the number of residues. Each point represents a protein in the calibration set of 104 proteins. Energies were calculated using the initial force field before parameter dynamics.

This resulted in the set of 104 proteins given in Table II. The choice of proteins was somewhat arbitrary, but generally erred on the side of reliable structures. This may have unfairly excluded high-quality structures, but allowed such structures to be used for testing.

Initial Parameter Estimation

The optimization of σ parameters for Equations (2) to (4) was a large calculation based on energy differences between native and misfolded structures. Initial σ values, however, were calculated using a simple method based only on native structures and designed to minimize the energy after summing across all proteins.[19] For example, for the pseudo-angle term, Equation (2), the derivative of the energy with respect to $\sigma_J^{n,n+2}$ was set to zero and solved for for $\sigma_J^{n,n+2}$, assuming $\epsilon_J^{n,n+2} = 1$. This yielded an expression.

$$\sigma_J^{n,n+2} = \left[\frac{\sum^{N_{J_{ijk}}} r_{ik}^{-q}}{\sum^{N_{J_{ijk}}} r_{ik}^{-12}}\right]^{1/(12-q)} \quad (5)$$

TABLE II. Proteins Used for Force Field Calibration

Protein code[a]	Number of subunits	Number of residues	Residue coordinates unknown[b]	Resolution (Å)	R factor[c]
1aaj		105		1.8	0.16
1aak		150		2.4	0.22
1aap	2	112		1.5	0.18
1aep		161	8NC	2.7	0.21
1apa		266	5N	2.3	0.17
1arb		268	5C	1.2	0.15
1ayh		214		2.0	0.16
1baa		243		2.8	0.20
1bgh		85		1.8	0.19
1bn2		354	87–95	2.8	0.23
1bov	5	345		2.2	0.18
1bsa	3	321		2.0	0.17
1cau	2	365		2.3	0.19
1cd8		114		2.6	0.19
1cdt	2	120		2.5	0.20
1cew		108		2.0	0.20
1cmb	2	208		1.8	0.20
1col	2	408	14NC	2.4	0.18
1cse	2	337		1.2	0.18
1ctf		68		1.7	0.17
1dfn	2	60		1.9	0.19
1dri		271		1.7	0.19
1dsb	2	376		2.0	0.17
1eaf		243		2.3	0.20
1ede		310		1.9	0.16
1end		137		1.6	0.20
1fas		61		1.8	0.15
1fba	4	1440		1.9	0.18
1gdh	2	640		2.4	0.19
1gmf	2	238		2.4	0.20
1hdd	2	114		2.8	0.22
1hle	2	375		2.0	0.18
1hsb	4	374		1.9	0.22
1hyp		75		1.8	0.19
1ifa		159		2.6	0.20
1ifc		131		1.2	0.17
1ipd		345		2.2	0.19
1lfb		77		2.8	0.21
1lis		131		1.9	0.19
1lts	7	741		2.0	0.18
1mpp		357		2.0	0.16
1nar		289		1.8	0.16
1ndk		148		2.2	0.20
1omp		370		1.8	0.21
1onc		104		1.7	0.18
1ppa		121		2.0	0.16
1ppb	3	298		1.9	0.16
1pgd		469		2.5	0.19
1pii		452		2.0	0.17
1plc		99		1.3	0.15
1poh		85		2.0	0.14
1pos	2	212		2.6	0.22
1rcb		129		2.2	0.22
1rtc		268		2.3	0.23
1rve	2	488		2.5	0.19
1sbp		309		1.7	0.18
1sgt		223		1.7	0.16
1shg		57		1.8	0.20
1shf	2	118		1.9	0.18

TABLE II. (Continued)

Protein code[a]	Number of subunits	Number of residues	Residue coordinates unknown[b]	Resolution (Å)	R factor[c]
1sim		381		2.0	0.19
1sry	2	842		2.5	0.18
1tbp	2	360		2.6	0.21
1tlk		103		2.8	0.18
1tml		286		1.8	0.18
1tpk	3	264		2.4	0.18
1ttb	2	254		1.7	0.16
1ula		289		2.8	0.20
1utg		70		1.3	0.23
1vaa	3	381		2.3	0.17
2alp		196		1.7	0.13
2cas		548		3.0	0.21
2cpl		164		1.6	0.18
2cro		65		2.4	0.19
2hpr		87		2.0	0.15
2liv		344		2.4	0.18
2pmg	2	1122		2.7	0.22
2pol	2	732		2.5	0.18
2rn2		155		1.5	0.20
2sas		185		2.4	0.20
2scp	2	348		2.0	0.18
2sga		181		1.5	0.13
2snv		151		2.8	0.20
2tgi		112		1.8	0.17
2zta	2	62		1.8	0.18
3adk		194		2.1	0.19
3cd4		178		2.2	0.20
3chy		128		1.7	0.15
3dpa		218		2.5	0.18
3eca	8	1308		2.4	0.15
3il8		68		2.0	0.19
3mon	8	376		2.8	0.19
3rp2	2	448		1.9	0.19
3sgb	2	235		1.8	0.12
3tgl		265		1.9	0.13
4blm	2	512		2.0	0.02
4enl		436		1.9	0.15
4gcr		174		1.5	0.18
4ins	4	102		1.5	0.15
4sgb	2	236		2.1	0.14
6taa		478	2C	2.1	0.20
7icd		414		2.4	0.18
8ilb		146		2.4	0.16
9rnt		104		1.5	0.14
9wga	2	342		1.8	0.17

[a] PDB acquisition code.

[b] Number of residues at the N or C terminus for which no coordinates are available.

[c] Crystallographic R factor.

where $q = 10$ for the pseudoangle term of Equation (2). The summation runs over all $N_{J_{ijk}}$ triplets of sequential residues not spanning chain breaks and across the protein calibration set and where the central residue is of amino acid class J. The expression is still valid if class J contains more than one amino acid type. Analogous expressions with $q = 10$ and $q = 1$ were used for the $\sigma_{JK}^{n,n+3}$ and $\sigma_{IJ}^{\text{long}}$ parameters in the pseudotorsion and long-range interaction terms, respectively.

The various ϵ parameters were calculated using a method designed to weight those interactions which showed a preference for a certain distance. Again, one can use the example of the parameter for the pseudoangle term as an example. If a certain residue type X causes a very narrow range of n, $n + 2$

distances, then the corresponding $\epsilon_X^{n,n+2}$ should be large. If residues surrounding residues of class Y occupy a wide range of distances, then the $\epsilon_Y^{n,n+2}$ parameter should be small. This statistical tendency can be quantified by first setting $\epsilon_J^{n,n+2} = 1$ in Equation (2) for the amino acid type class J and then summing over all relevant pairs to obtain

$$\epsilon_{\mathrm{J}}^{n,n+2} = \frac{\sum^{N_{J_{ijk}}} E_{J_{ijk}}^{n,n,+2}}{N_{J_{ijk}}}\,. \tag{6}$$

As for the previous calculations, the summation runs over all triplets of adjacent residues not spanning a chain break and where the central residue is of amino acid class J. The same expression, summed over the appropriate pair energies was used for the n, $n + 3$ and long-range interactions to generate $\epsilon^{n,n+3}$ and ϵ^{long} parameters, respectively.

The example in Equation (6) is based on the n, $n + 2$ term, but a physical basis for the method can be seen by considering the example of the long-range interaction parameters. A pair of residues such as Asp and Arg are oppositely charged and may have a strong statistical preference for interacting at a certain distance (close to $\sigma_{\mathrm{Asp\text{-}Arg}}^{\mathrm{long}}$). This distance of minimum interaction energy will result in a maximum value for Equation (6). By contrast, Gly-Gly pairs show little tendency to prefer any particular distance. Energies calculated from Equation (4) will often be near zero, and Equation (6) will yield a value closer to zero. This interaction will have correspondingly little weight in the final force field.

Parameter Classification

For the pseudoangle term, Equation (2), the choice of parameter depended on the type of the central residue, so there were 20 $\sigma^{n,n+2}$ and 20 $\epsilon^{n,n+2}$ parameters or, in this nomenclature, 20 amino acid classes, each having one member. No parameter reduction was applied to these terms.

For the n, $n + 3$ and long-range parameters, an algorithm was used to reduce the number of adjustable parameters and improve the method's ability to generalize.[18] Both of these potential energy functions depended on the type of two residues, so without some classification, there would be $(20 \times 21)/2 = 210$ pairwise interaction parameters.

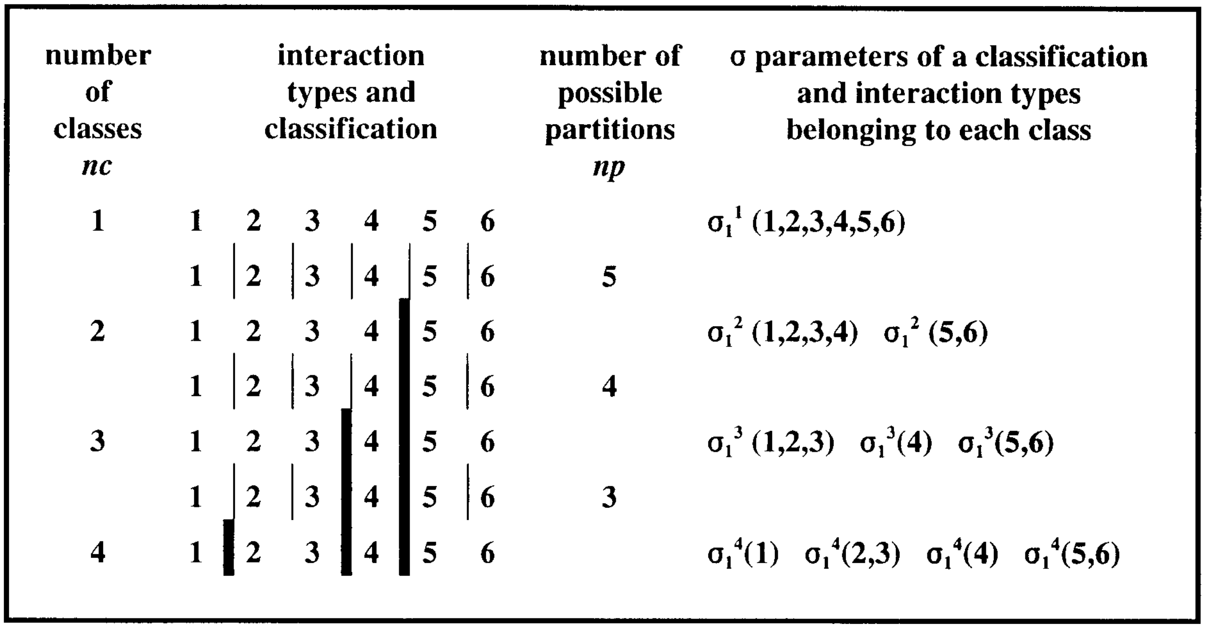
The aim of the algorithm was to produce a classification that reduced the total energy summed across the reference protein set. Physically, this could be seen as looking for those interactions that were most similar and putting them together in a class.

The method was applied separately for the n, $n + 3$ and long-range interaction terms and required that for any group or class of pairwise interactions, one could use Equation (5) to calculate a σ (distance of lowest energy) for that group or class. For example, a class may consist of Asp-Cys and Glu-Leu interactions. Then, for all interactions of these types across the calibration set of 104 proteins, a σ could be calculated. Given some classification (indicated by C^n), one can calculate a total energy over all interaction classes and all proteins. This is the energy $E(C^n)$ associated with some classification, C^n.

For initialization, one treats each interaction type separately and calculates a σ value for each pairwise interaction type using Equation (5). This forms a list that can be sorted according to σ value. The result will be that pairs that tend to be close to each other (small σ) are at one end of the list and pairs that repel (in a statistical sense) will be at the other end of the list.

The classification procedure now operates by initially putting all interactions into a single class. So, all members of the list form class 1 (indicated by the lower index), C_1^1, of the classification comprising only one class (indicated by the upper index, nc). When calculating the σ value σ_1^1 for class C_1^1, the summation in Equation (5) runs over all pairwise interaction types of class C_1^1. At each step n (≥ 2) of the classification procedure, one of the classes present is partitioned into two parts such that the decrease in the energy of classification C^n, $E(C^n)$, with respect to the energy $E(C^{n-1})$ is maximized. This can be illustrated by an example. Figure 2 shows a simple case with six σ values, each corresponding to a different pairwise interaction type. The target in this example was arbitrarily set to 4 classes ($nt = 4$). Initial estimates of σ were calculated for each interaction type, sorted and put into one class ($nc = 1$), as shown in the top row of the figure. The σ value of this class, σ_1^1 depends on the summation over all six interaction types in Equation (5). The second row shows $np = 5$ possible places to divide the set, each marked by a light line. Each position is tested in turn and the one which produces the lowest energy selected. This results in two classes (third row) where the first class consists of interactions types 1 to 4 and the second of interaction types 5 and 6. There are then $np = 4$ possible ways to add a division (fourth row). Picking the position that results in lowest energy splits the first class. The result is shown on the fifth row with three classes of σ values. At the next step, there are $np = 3$ possible places to divide a class. Choosing one of these results in the final row with four classes. Since the number of classes equals the target number, ($nc = nt$), the algorithm stops.

The algorithm can now be described more formally. Let C be some set (class) of pairwise interaction types. Initially, there is one set, C_1 containing all pairwise interaction types (210 in this calculation). By the end of the procedure there will be sets of interaction types $C_1 \ldots C_{nt}$ forming a classification C^{nt}. The classification or set of interaction classes is denoted S. Because one needs to rank classifications in terms of energy, the algorithm also requires a set



number of classes nc	interaction types and classification	number of possible partitions np	σ parameters of a classification and interaction types belonging to each class
1	1 2 3 4 5 6		σ_1^1 (1,2,3,4,5,6)
	1 | 2 | 3 | 4 | 5 | 6	5	
2	1 2 3 4 ▌5 6		σ_1^2 (1,2,3,4) σ_1^2 (5,6)
	1 | 2 | 3 | 4 ▌5 | 6	4	
3	1 2 3 ▌4 ▌5 6		σ_1^3 (1,2,3) σ_1^3(4) σ_1^3(5,6)
	1 | 2 | 3 ▌4 ▌5 | 6	3	
4	1 ▌2 3 ▌4 ▌5 6		σ_1^4(1) σ_1^4(2,3) σ_1^4(4) σ_1^4(5,6)

Fig. 2. Algorithm for classification of pairwise interaction types. Each number (1–6) represents a pairwise interaction. Subscripts indicate the class to which the σ is assigned. Thin lines show possible divisions of the set. Thick lines show divisions introduced during the progress of the algorithm. Steps are described in text.

K where each member is an energy, calculated over the whole set of calibration proteins. Each energy, E_{Dy} will correspond to a trial classification C_{Dy}^{nc+1}, which is obtained from classification C^{nc} by partitioning class D (or C_D^{nc}) of C^{nc} into two parts indicated by y. So, one has $E_{Dy} = E(C_{Dy}^{nc+1})$. For a pseudocode description we use a generic SET data type:

```
variables
$C_1 \ldots C_{nt}$   SET    /* Each set is a class of interactions */
$S$                   SET    /*Classification $C^{nc}$, a set of $C_1 \ldots C_{nc}$ */
$K$                   SET    /* Set of energies */
$L, L_1, L_2$         SET    /* Set of interactions */
pseudocode
$C_1 := \{x | x$ is a pairwise interaction type$\}$
sort $C$ according to the $\sigma$ for each interaction type as calculated through Equation (5)
$S := \{C_1\}$        /* The first member of the set of sets is $C_1$ */
$nc := 1$
while ($nc < nt$) {
  $K := \varnothing$          /* Initialize to empty set */
  for ($D :=$ each existing class $C$) {
    for ($y :=$ every possible division of class $D$){
      $E_{Dy} :=$ total energy associated with class $D$ and division $y$
      $K := K \cup \{E_{Dy}\}$          /* Add this energy to set $K$ */
    }
  }
  sort set $K$ according to $E_{Dy}$     /* find the division giving lowest total energy */
  let $L$ be the set associated with the lowest energy member of $K$
  split $L$ into components $L_1$ and $L_2$ according to the division $y$ referenced by $E_{Dy}$
  $S := S - L$          /* remove the set which is to be split */
  $S := S \cup L_1$     /* Add first component of new set */
  $S := S \cup L_2$     /* Add second component of new set */
  $nc := nc + 1$
```

This scheme was used to reduce the number of n, $n + 3$ interaction pair types from 210 to 36 ($nt = 36$) and the long-range interaction pair types from 210 to 10 ($nt = 10$). This was the same number of classes as in early work by Crippen and Snow,[19] but the classes here were for interaction types rather than amino acid types. This means, for example, that a Trp-Tyr interaction may be in one class, while a Trp-Gly interaction could be in another.

Generation and Initial Selection of Alternative Structures

Compact misfolded alternative structures were generated for each of the 104 native structures in the calibration set. These alternative structures were constructed by taking each native sequence and threading it on to the C^α coordinates of every other protein with at least the same number of residues.[24,25] If the native structure had N_{nat} residues, then the first alternative structure used residues 1 to N_{nat} in the first larger (template) structure. The second alternative structure came from the C^α coordinates of residues 2 to $N_{\mathrm{nat}} + 1$ in the template structure and so on. Thus, each template of N_{tmpl} residues added $N_{\mathrm{tmpl}} - N_{\mathrm{nat}} + 1$ alternative conformations. Unlike in previous work, the number of alternative structures was then doubled by threading each sequence backward on template coordinates. Lastly, each native structure provided one more alternative structure by threading the sequence backward onto its own native coordinates. As templates, only unbroken protein subunits were used. Proteins with chain breaks were threaded on to larger template structures with a 1-residue gap at the chain break.

Some alternative structures were then discarded. An alternative structure, which was structurally similar to a native, was not used. Ideally, redundancy in the alternative structures would be removed by comparing every alternative conformation with every other alternative conformation, but this would be a very expensive calculation for little gain, since the calibration set of 104 proteins was chosen so as to have little internal homology. A more practical approach was to compare each alternative conformation with the previously generated alternative conformation and only accept it if it was significantly different. This procedure was very efficient, since successively generated alternative structures are most likely to be similar to each other.

Structural similarity was defined by comparing distance matrices from structures. This measure has been called the distance matrix error (DME)[26] or distance root mean square difference (D_{rmsd})[27] and is given by

$$D_{\mathrm{rmsd}}(A, B) = \left[\frac{2}{N_{\mathrm{res}}(N_{\mathrm{res}} - 1)} \sum_{i<j}^{N_{\mathrm{res}}} (r_{Aij} - r_{Bij})^2\right]^{1/2} \qquad (7)$$

where N_{res} is the number of residues, and r_{Aij} and r_{Bij} are the distances between particles i and j in structures A and B, respectively. The threshold for structural similarity, D, has a cube root dependence on the number of amino acids. We used the value from Maiorov and Crippen[27]

$$D^{\sigma} = -4.54 + 2.36\,(N_{\mathrm{res}})^{1/3} \qquad (8)$$

Although their prescription was based on the root mean square difference of cartesian coordinates, the effect in this work was that too low a value for D^{sig} was used and the discrimination problem (below) was made more difficult.

Parameter Dynamics and σ Optimization

The initial parameter estimates, described above, were based on native protein structures only. In the next step, parameters were optimized so as to simultaneously disfavor incorrectly folded alternative structures. This was done using an approximation to newtonian dynamics in parameter space. The word approximation is deliberately chosen, since the method did not conserve (parameter) energy, the parameter trajectories did not have continuous first derivatives, and the scheme relied on a series of pragmatic decisions necessary to make the calculations computationally tractable.

The ϵ parameters [Eqs. (2) to (4)] were not optimized. In our formulation, they are scaling factors and so attempting to optimize them results in infinite values. Only σ values for the $(n, n + 2)$, $(n,n + 3)$ and long-range terms were subjected to dynamics.

The potential energy function for parameters consisted of three terms. The first two were based on the goals of the parametrization. First, native structures should be of low energy. Second, misfolded alternatives should be of higher energy. Before writing this formally, one should note that the number of alternative structures per protein differs within the protein calibration set and the proteins themselves differ in the number of interactions they contribute. This means that a large protein (few alternative structures) would dominate the native structure term, while a small protein would dominate the alternative structure term. The scheme here first normalized to account for the number of alternative structures for a protein and then scaled the total parameter force due to each protein to be of the same magnitude. The parameter energy is of a different form to the pseudoenergies calculated for protein structures, so it is denoted E^{σ}. For a single protein and its alternative structures, E^{σ} is given by

$$E^{\sigma}(\mathbf{r}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}) = E_{\mathrm{total}}(\mathbf{r}^{\mathrm{NAT}}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}) - \frac{1}{N_{\mathrm{alt}}} \sum_{\alpha=1}^{N_{\mathrm{alt}}} E_{\mathrm{total}}(\mathbf{r}^{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}) + \frac{1}{N_{\mathrm{prot}}} E_{\mathrm{restr}}(\sigma) \qquad (9)$$

where $\mathbf{r}^{\mathrm{NAT}}$ is the coordinate vector for the native structure and N_{prot} is the total number of proteins in the calibration set. The summation in the second term runs over the pseudoenergy of each of the N_{alt} alternative structures, with coordinates labeled $\mathbf{r}^{\alpha}$. $\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma}$ are the parameter vectors as used in Equations (2) to (4). E_{restr} was a device to dissuade individual parameters σ of the vector $\boldsymbol{\sigma}$ from entering unlikely areas of parameter space and consisted of harmonic terms to enforce minimum and maximum σ values (σ_{min} and σ_{max}, respectively), given by

$$E_{\mathrm{restr}}(\sigma) \begin{cases} = \dfrac{k_{\mathrm{restr}}}{2}(\sigma - \sigma_{\mathrm{min}})^2 & \text{if} \quad \sigma < \sigma_{\mathrm{min}} \\ = \dfrac{k_{\mathrm{restr}}}{2}(\sigma - \sigma_{\mathrm{max}})^2 & \text{if} \quad \sigma < \sigma_{\mathrm{max}} \end{cases} \qquad (10)$$

k_{restr} was set to 200 energy units/Å^2 and σ_{min} to 3 Å. σ_{max} was set to two pseudobond lengths ($2 \times 3.8 = 7.6$ Å) for the n, $n + 2$ pseudoenergy term [Eq. (2)] and three pseudobond lengths (11.4 Å) for the n, $n + 3$ pseudoenergy term [Eq. (3)].

An initial force, $\mathbf{F}_{\mathrm{ini}}$, acting on the elements of $\boldsymbol{\sigma}$, due to each protein and its alternative structures was calculated by taking the opposite of the derivative of Equation (9) with respect to $\boldsymbol{\sigma}$. The final force vector, $\mathbf{F}_{\text{total}}$ was given by

$$\mathbf{F}_{\text{total}}(\mathbf{r}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}) = \frac{1}{N_{\text{prot}}} \sum_{I=1}^{N_{\text{prot}}} \left(\frac{\mathbf{F}^{I}_{\text{ini}}(\mathbf{r}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}}{|\mathbf{F}^{I}_{\text{ini}}\,(\mathbf{r}, \boldsymbol{\epsilon}, \boldsymbol{\sigma}|} \right) \qquad (11)$$

where F^{1}_{ini} is the force due to protein I and its alternative structures and the summation runs over all N_{prot} proteins in the calibration set. Forcing the contribution of each protein to be the same does not lead to discontinuities in the parameter trajectories, but potentially violates energy conservation.

Parameter dynamics were run with arbitrary units. Boltzmann's constant was taken as 1, and, by trial and error, a time step of 5×10^{-4} was chosen for integrating the equations of motion. Initial temperatures were not taken from a maxwellian distribution, as is common practice in molecular dynamics simulations. A less rigorous approach was used where parameters were allowed to develop velocities by moving in the field of Equation (9) and the temperature maintained by coupling to a heat bath.[28] Separate temperature baths were kept for σ^{long} parameters [Eq. (4)] and for the $\sigma^{n,n+2}$ and $\sigma^{n,n+3}$ parameters [Eqs. (2) and (3)]. The baths were set at 20 and 70 temperature units, respectively. Coupling to both baths was initially set to 5×10^{-3} time units and this was increased by an exponential scheme to 7.1×10^{-3} over 100 time steps. At each time step, the SHAKE algorithm[21] was applied to all native structures and their alternatives to bring pseudobonds to regular lengths.

The scheme described was then modified to improve its ability to discriminate correctly from incorrectly folded structures and to speed the calculations. The initial parameters showed some capability for discrimination. That is, many misfolded structures were of higher energy than native conformations, and trial calculations suggested that they remain of high energy after parameter dynamics. Clearly, they contributed little of use to the force experienced by the parameters. This led to a screening, applied every step of parameter dynamics, to select misfolded alternative structures of low energy. Intuitively, one might choose alternative structures of energy lower than the corresponding native structure, but this would not be a strong enough condition for a useful force field. One wants the difference between correct and incorrect structures to be as large as possible. To this end, a threshold energy (E_{thresh}) was defined, which should be more positive than the native energy by some positive number Δ. We then define

$$E^{I}_{\text{thresh}} = E^{I}_{\text{nat}} + \Delta^{I} \qquad (12)$$

where the superscript I, denotes protein I, so E^{I}_{nat} is the energy of native structure I and E^{I}_{thresh} is the threshold energy for protein I; An alternative conformation α will only be used in the force calculation if its energy E^{I}_{α} is below E^{I}_{thres}. From this point, one can see that Δ^{I} should be gradually increased as the force field improves. This idea is shown in Figure 3. Each point in the figure represents an alternative structure. The y-axis gives the energy corresponding to each conformation. Points marked by dots are alternative structures of high energy and do not contribute to the force felt by parameters. Conformations marked by crosses are alternative structures of low energy, which should be used to calculate forces on parameters. Initially (Fig. 3a), there are three conformations below the energy threshold. After some steps of parameter dynamics the force field has improved and the energy of two of the alternative structures rises. In Figure 3b, one of the three crossed points is above the energy threshold. The threshold should be raised as shown in Figure 3c, so as to select the most important (lowest energy) alternative structures. One still requires a method for determining the way in which the threshold difference Δ^{I} is increased.

To this end, weak coupling[28] was used. First, one needs a switching function, to determine whether or not an alternative structure, α, is below the threshold relative to the native structure of protein I. This is defined by

$$s(E^{I}_{\text{nat}}, E^{I}_{\alpha}, \Delta^{I}) = \begin{cases} 0 & \text{if} \quad E^{I}_{\alpha} > E^{I}_{\text{nat}} + \Delta^{I} \\ 1 & \text{if} \quad E^{I}_{\alpha} \leq E^{I}_{\text{nat}} + \Delta^{I} \end{cases}. \qquad (13)$$

This switching function can then be used to define the average energy difference between low-energy alternative structures and the threshold. If there are N_{prot} proteins in the calibration set and Δ_{av} is the average difference between threshold energies E^{I}_{thresh} and the energies of alternative structures E^{I}_{a}, one can write

$$\Delta_{\text{av}} = \frac{1}{N_{\text{prot}}} \sum_{I=1}^{N_{\text{prot}}} \cdot \left(\frac{\sum_{\alpha=1}^{N^{I}_{\text{alt}}} [(E^{I}_{\text{nat}} + \Delta^{I} - E^{I}_{\alpha}) s(E^{I}_{\text{nat}}, E^{I}_{\alpha}, \Delta^{I})]}{\sum \Sigma^{N^{I}_{\text{alt}}}_{\alpha=1} s(E^{I}_{\text{nat}}, E^{I}_{\alpha}, \Delta^{I})} \right), \qquad (14)$$

To continue the analogy with weak coupling schemes,[28] a target or reference value, Δ^{0}_{av}, was set, and the goal was to maintain this at a constant value. A scaling factor, λ, was calculated from

$$\lambda = \left[1 + \frac{\delta t}{\tau} \left(\frac{\Delta^{0}_{\text{av}}}{\Delta_{\text{av}}} - 1 \right) \right]^{1/2} \qquad (15)$$

where δt is the time step of the integrator for the parameter dynamics, and τ is the coupling constant that controls how rapidly changes are applied. Fi-

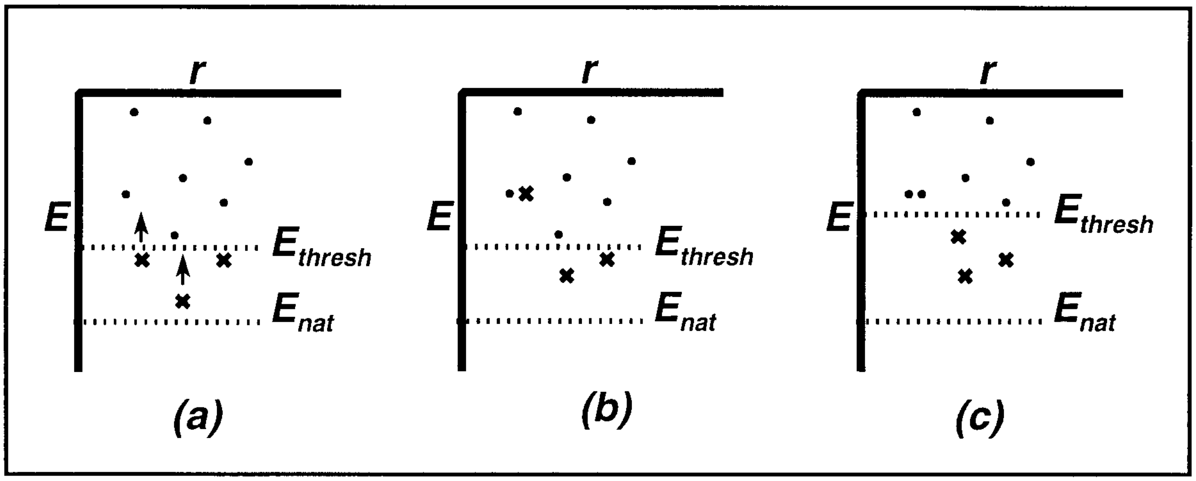


Fig. 3. Energy threshold for selection of alternative structures. *r* is some generalized coordinate representing conformational space. *E* denotes conformational energy. E_{nat} is the energy of a native structure and E_{thresh} the corresponding threshold energy for alternative conformations. Crosses mark conformations of energy less than the threshold that contribute to the calculation of forces on the parameters. Dots mark the alternatives which are already of high energy and do not contribute to the force calculation.

nally, the coupling was applied by applying the scaling to the energy difference Δ^I:

$$\Delta^I(t) = \lambda\Delta^I(t - \delta t) \qquad (16)$$

where $\Delta^I(t)$ is the value of Δ at the current time step and $\Delta^I(t - \delta t)$ is the value of Δ^1 at the previous time step.

The constants in Equation (15) were set by trial and error. The fraction $\delta t/\tau$ was set to 0.1 and Δ^0_{av} to 10.0. One should also note that $\Delta^I(t)$ was different for each protein *I*. Furthermore, $\Delta^I(t)$ is not defined at the first time step, so we arbitrarily set $\Delta^I(0) = |E^I|$, the absolute value of the energy of the native structure of protein *I*. In practice, this value was probably too large. During the short parameter dynamics calculation, the system did not equilibrate with the weak coupling bath, $\Delta^I(t)$ tended to decrease, and the number of alternative structures contributing to Figure 3 the parameter forces decreased correspondingly.

RESULTS

The first calculation was the classification of parameters for the *n*, *n* + 3 and long-range interactions, given in Tables III and IV, respectively. The classifications were based on the initial estimates of $\sigma^{n,n+3}$ and σ^{long} before parameter dynamics and do not include the influence of misfolded alternative structures. The classes were ordered by the initial σ value and show some trends. For example, the glycine column for the long-range interaction (Table IV) shows that the residue is usually in classes 1, 2, or 3, favoring the smallest distance of minimum interaction energy. Similarly, the bulky aliphatic leucine residue is usually in classes 6–10 with a tendency to prefer interactions at larger distances. While one should not read too much physical significance into this initial calculation, one property is clear. Classifying interaction types rather than amino acids allows more flexibility in the functional form without increasing the number of adjustable parameters. For example, one would expect the oppositely charged Asp and Arg to be in different classes, but the scheme here allows Asp and Arg to fall into the same interaction class with respect to residues such as Trp or Gly.

The 104 proteins in the calibration set resulted in 1.7 × 10^6 misfolded alternative structures. After applying similarity criteria and the initial selection based on energies, there remained 350,000 alternative structures (summed over all 104 proteins). This set formed the basis for calculating forces on parameters during dynamics calculations, although all 1.7 × 10^6 alternative structures were used for testing force field quality below. The example dynamics calculation shown here ran for (coincidentally) 104 steps. The best results in terms of discriminating correct and misfolded structures were obtained after 92 steps and are described below.

The results of the parameter dynamics calculations can be most clearly seen by considering the best and worst results. A good result would be one where the native structure is of low energy compared to all alternative structures. Using this measure, the protein giving the best result is a 60-residue toxin with the PDB acquisition code, 1cdt shown in Figure 4. The top panel shows the energy of alternative structures with respect to the energy of the native structure. Energies are divided by the number of residues

TABLE III. ***n,n*** **+ 3 Interaction Pair Classes Identified by Their Integer Class Number (1–36)**

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	14	14	19	19	11	15	18	20	15	1	14	16	19	22	19	22	10	16	14	13
Arg	14	10	11	17	13	14	15	29	9	14	12	9	18	12	29	23	14	22	19	8
Asn	19	11	33	15	20	17	20	23	19	12	21	23	27	25	6	28	27	24	20	20
Asp	19	17	15	24	24	14	13	26	20	10	14	23	19	17	5	24	21	25	16	15
Cys	11	13	20	24	22	13	19	34	17	24	18	27	6	25	32	26	30	30	15	29
Gln	15	14	17	14	13	9	16	27	18	12	19	21	11	23	3	17	10	9	14	23
Glu	18	15	20	13	19	16	14	25	22	14	19	18	14	7	19	16	19	24	17	13
Gly	20	29	23	26	34	27	25	28	30	2	20	14	16	12	29	29	29	29	27	25
His	15	9	19	20	17	18	22	30	25	25	20	21	28	13	31	23	22	10	9	14
Ile	1	14	12	10	24	12	14	2	25	12	16	9	19	14	19	21	24	22	13	14
Leu	14	12	21	14	18	19	19	20	20	16	18	19	11	9	25	17	21	9	19	16
Lys	16	9	23	23	27	21	18	14	21	9	19	12	21	20	28	18	22	12	13	13
Met	19	18	27	19	6	11	14	16	28	19	11	21	10	20	22	20	10	17	27	16
Phe	22	12	25	17	25	23	7	12	13	14	9	20	20	12	30	29	9	8	15	11
Pro	19	29	6	5	32	3	19	29	31	19	25	28	22	30	35	6	14	13	28	22
Ser	22	23	28	24	26	17	16	29	23	21	17	18	20	29	6	24	22	29	10	25
Thr	10	14	27	21	30	10	19	29	22	24	21	22	10	9	14	22	27	4	4	20
Trp	16	22	24	25	30	9	24	29	10	22	9	12	17	8	13	29	4	36	7	8
Tyr	14	19	20	16	15	14	17	27	9	13	19	13	27	15	28	10	4	7	11	22
Val	13	8	20	15	29	23	13	25	14	14	16	13	16	11	22	25	20	8	22	8

TABLE IV. Long-Range Interaction Pair Classes Identified by Their Integer Class Number (1–20)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	The	Trp	Tyr	Val
Ala	5	7	8	5	4	7	6	2	9	4	6	7	3	6	4	2	5	6	5	6
Arg	7	5	5	4	4	8	5	2	5	5	8	9	8	8	4	4	3	5	4	7
Asn	8	5	4	7	7	8	6	2	8	8	9	3	8	5	3	6	7	8	5	8
Asp	5	4	7	8	10	8	9	2	4	9	9	3	6	4	7	6	4	8	9	9
Cys	4	4	7	10	1	8	9	2	9	4	6	8	3	2	4	7	3	3	3	6
Gln	7	8	8	8	8	4	8	2	6	8	9	8	8	6	6	3	3	7	8	7
Glu	6	5	6	9	9	8	6	4	8	7	10	3	7	7	9	5	5	9	7	8
Gly	2	2	2	2	2	2	4	1	3	3	3	3	3	2	2	2	2	2	2	3
His	9	5	8	4	9	6	8	3	3	8	9	3	6	4	7	7	6	2	6	9
Ile	4	5	8	9	4	8	7	3	8	5	6	7	4	7	8	7	7	6	4	4
Leu	6	8	9	9	6	9	10	3	9	6	8	9	6	7	8	7	9	6	6	6
Lys	7	9	3	3	8	8	3	3	3	7	9	3	9	7	3	4	3	3	3	8
Met	3	8	8	6	3	8	7	3	6	4	6	9	9	5	3	7	7	8	7	7
Phe	6	8	5	4	2	6	7	2	4	7	7	7	5	3	5	5	4	6	8	4
Pro	4	4	3	7	4	6	9	2	7	8	8	3	3	5	2	2	5	2	5	6
Ser	2	4	6	6	7	3	5	2	7	7	7	4	7	5	2	3	3	3	3	4
Thr	5	3	7	4	3	3	5	2	6	7	9	3	7	4	5	3	2	4	4	4
Trp	6	5	8	8	3	7	9	2	2	6	6	3	8	6	2	3	4	2	2	6
Tyr	5	4	5	9	3	8	7	2	6	4	6	3	7	8	5	3	4	2	5	5
Val	6	7	8	9	6	7	8	3	9	4	6	8	7	4	6	4	4	6	5	4

as this allows an approximate comparison of proteins. It is encouraging that the alternative structures are clearly of higher energy and the progress of the dynamics calculation is shown in the middle panel. This gives the number of alternative structures below the energy threshold [Eq. (12)] at each time step. Although initially there are nearly 6,000 alternatives of low energy, the parameters are very quickly driven to favor the native structure. The bottom panel gives some insight into the path of the calculation, showing how the σ^{long} parameters change during the calculation.

Using the same criterion for quality, the worst results came from the protein 1ppa and are shown in Figure 5. The top panel again shows the energy of alternative structures relative to the energy of the native (after dividing by the number of residues). More than half the alternative conformations are of lower energy. There is no reason to suspect a problem with native coordinates, but the middle panel may give some clue as to why the parametrization has been so unsuccessful for this protein. From around step 45, the number of alternative structures below the threshold actually increases steadily. The most likely explanation is that the parameter trajectories are being driven most strongly by the energy of other proteins in the calibration set, unfortunately, at the expense of 1ppa. It is also worth examining the

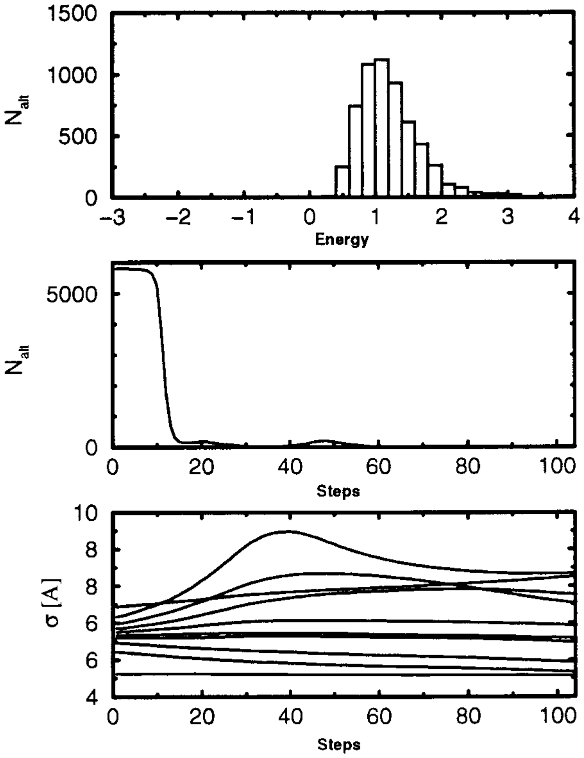


Fig. 4. Progress of parametrization for protein 1cdt. **Top:** Histogram showing the energy distribution of alternative structures for 1cdt. Energies are relative to the native structure ($E_{alt} - E_{nat}$) and are divided by the number of residues in the native protein. Energies were calculated with parameters after 92 steps of dynamics. **Middle:** Number of alternative structures of low energy (below the energy threshold) at each time step of the parameter dynamics calculation. **Bottom** Value of the first 10 σ^{long} parameters (Table VII) at each time step in the parameter dynamics calculation.

bottom panel, again showing the σ^{long} parameters as a function of time. After 104 steps, they have not stopped fluctuating. If one was to pursue parameters for this particular force field parametrization, a longer calculation would be necessary.

An overview of the results is given in Figure 6. This shows a comparison of the energies of alternative structures (relative to the native energy) for 1cdt, 1ppa, and the calibration set average. This last histogram is obtained by summing over all native and alternative structures and dividing the ordinate by the number of proteins (104) and always dividing the energy difference ($E_{alt} - E_{nat}$) by the number of residues in the native structure. In an approximate statistical sense, the simple force field is quite powerful and usually able to discriminate correct (native) structures from plausible, but incorrect, structures. Unfortunately, it could not be deemed truly predictive. For an application such as threading (testing structures for a sequence), this particular parametrization would yield too many false positives to be useful.

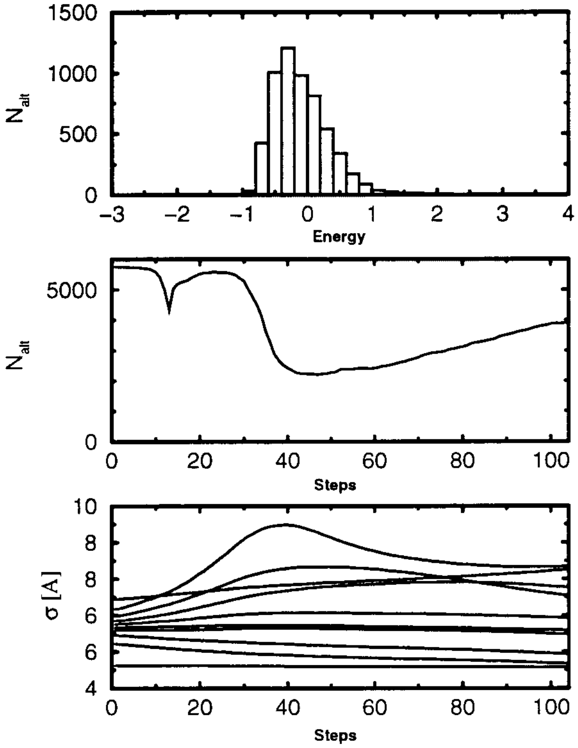


Fig. 5. Progress of parametrization for protein 1ppa. Panels correspond to Figure 4.

The final results of the parametrization calculations are given in Tables V, VI, and VII. When reading Tables V and VI it is important to remember the amino acids and pairs of amino acids are the central ones to each interaction. For example, the first line of Table VI refers to Ala-Ile, so this parameter pertains to alanine and isoleucine as the *j*, *k* pair of an *i*, *j*, *k*, *l* quartet. Viewed from a historical perspective, this is similar to formalizing the tendency of amino acids to be in certain secondary structures and quantifying this geometrically. At the same time, interpretation in terms of secondary structure is not so straightforward. The n, $n + 2$ and n, $n + 3$ parameters are determined in the field due to all the proteins, which, in turn, depends on all the parameters. This means the parameters reflecting short-range interactions may be dominated by secondary structure, but would not be sufficient to predict it. Furthermore, there is a more subtle restriction on their use. The initial parameters reflect average geometric considerations from Equation (5), but the final parameters reflect the influence of misfolded structures. For example, the $\sigma_{Val}^{n,n+2}$ was initially estimated to be 5.6 Å suggesting that when a valine residue is at the center of a triplet of residues there is a tendency for the outer two residues to be at a relatively large distance. After the parameter dynamics, this value becomes 2.7 Å suggesting that the ability to distinguish correct from

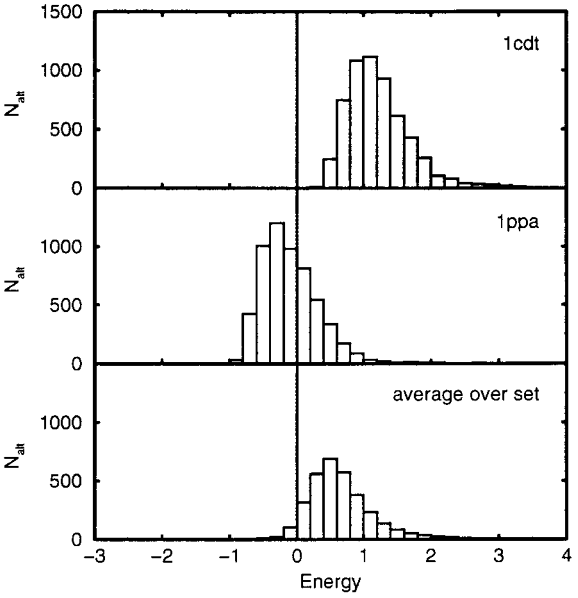


Fig. 6. Overall performance of parametrization. Each panel is a histogram showing the energy distribution of alternative structures relative to the native. Bars to the left of zero show alternative structures of energy lower than the corresponding native structure. **Top:** Results for 1cdt. **Middle:** Results for 1ppa. **Bottom:** Results summed over the 104 proteins of the calibration set and then divided by the number of proteins.

TABLE V. *n*,*n* + 2 Interaction Parameters

	$\sigma^{n,n+2}$ (Å)		$\epsilon^{n,n+2}$		
Class	Initial	Final	(energy)[a]	Member	N_{occur}[b]
1	5.470	5.249	1.460	Ala	2480
2	5.501	4.883	1.332	Arg	1356
3	5.535	5.170	1.454	Asn	1465
4	5.524	5.251	1.492	Asp	1677
5	5.679	6.225	1.194	Cys	595
6	5.461	5.002	1.412	Gln	1098
7	5.480	5.199	1.464	Glu	1762
8	5.581	4.136	1.256	Gly	2385
9	5.589	5.777	1.306	His	505
10	5.569	5.423	1.250	Ile	1581
11	5.546	5.731	1.406	Leu	2354
12	5.495	4.412	1.342	Lys	1830
13	5.513	5.107	1.304	Met	559
14	5.590	5.216	1.200	Phe	1141
15	5.259	4.235	0.960	Pro	1344
16	5.516	4.704	1.240	Ser	1811
17	5.626	3.948	1.220	Thr	1833
18	5.558	5.613	1.348	Trp	380
19	5.612	6.267	1.224	Tyr	1125
20	5.633	2.692	1.248	Val	1977

[a] Arbitrary energy units.

[b] Number of occurrences of this interaction in the calibration protein set.

incorrect structures is enhanced by preferring a short distance for this interaction.

The tables also contain information about the relative weights of the interactions since the ϵ parameters from Equations (2)–(4) directly scale the energy contribution of each term. For example, the $\sigma^{n,n+3}_{\text{Gly-Ile}}$ parameter given in Table VI is one of the shorter distances, but $\sigma^{n,n+3}_{\text{Gly-Ile}}$ is the smallest in the table (0.091 energy units) meaning that it is also the least significant of the *n*, *n* + 3 interactions.

One can again look for physical significance in the results, bearing in mind that the influence of misfolded structures in the parameter dynamics tends to weaken the physical significance of the parameters. Considering the long-range interaction parameters in Table VII, the Cys-Cys and Gly-Gly interaction pairs show a preference for interactions at the shortest distances (5.1 Å). This obviously reflects disulfide bonds and the fact that glycine possesses the smallest side chain. Beyond that, one would only want to point to general trends such as the smaller side chains usually preferring short interaction distances and interactions with bulky side chains such as valine or leucine tending to prefer larger interaction distances.

So far, we have only considered the force field's performance in recall. That is, how well the force field works within the range of structures used to construct it. It is more interesting to look at the method's ability to generalize. This is a measure of its predictive ability for proteins not used in the parametrization calculations. Ideally, one would perform a jackknife test, leaving each protein out of the calculation and reserving it for testing. This would be far too expensive computationally, but it is still possible to test the force field outside of the calibration set. Choosing the proteins for calibration involved ranking the set according to quality (resolution and R factor). The 10 proteins immediately following the calibration set and some arbitrary examples were then of good quality and given the selection criteria, guaranteed to be nonhomologous with the proteins used for calibration. Table VIII gives the results for 10 structures under the columns labeled "Before minimization." The immediate impression is that the parameters do not do much better than a set of random numbers, and the results range from excellent to bizarre. For 4pti, the discrimination is excellent with 99.8% of alternative structures being of higher energy. For 1abp, 99.9% of alternative structures appear to be of lower energy than the native. The reason for this is clear from the results of other workers.[19,25] A structure may be very near in space to an energetic minimum, but of apparently high energy. The remedy for this is to compare structures after energy minimizing. This is shown by the two columns labeled "After minimization" in Table VIII. Native and alternative conformations were subjected to no more than 25 steps of conjugate gradients minimization. From these results, it is clear that the force field does have a real ability to discriminate correct from incorrectly folded

TABLE VI. *n,n* + 3 Interaction Parameters

	$\sigma^{n,n+3}$ (Å)		$\epsilon^{n,n+3}$			
Class	Initial	Final	(energy)[a]	N_{mem}[b]	N_{occur}[c]	Members
1	4.057	4.094	0.190	1	282	Ala-Ile
2	4.247	4.228	0.091	1	249	Gly-Ile
3	4.418	4.403	0.142	1	86	Gln-Pro
4	4.585	4.530	0.153	2	213	Thr-Trp, Thr-Tyr
5	4.715	4.703	0.250	1	148	Asp-Pro
6	4.777	4.601	0.227	3	352	Asn-Pro, Cys-Met, Pro-Ser
7	4.901	4.748	0.693	2	129	Glu-Phe, Trp-Tyr
8	4.958	4.816	0.565	4	373	Arg-Val, Phe-Trp, Trp-Val, Val-Val
9	4.988	4.626	0.646	9	935	Arg-His, Arg-Lys, Gln-Gln, Gln-Trp, His-Tyr, Ile-Lys, Leu-Phe, Leu-Trp, Phe-Thr
10	5.012	4.858	0.609	8	919	Ala-Thr, Arg-Arg, Asp-Ile, Gln-Thr, His-Trp, Met-Met, Met-Thr, Ser-Tyr
11	5.034	4.914	0.637	6	537	Ala-Cys,, Arg-Asn, Gln-Met, Leu-Met, Phe-Val, Tyr-Tyr
12	5.053	4.786	0.665	9	1154	Arg-Leu, Arg-Phe, Asn-Ile, Gln-Ile, Gly-Phe, Ile-Ile, Lys-Lys, Lys-Trp, Phe-Phe
13	5.070	4.681	0.683	10	1493	Ala-Val, Arg-Cys, Asp-Glu, Cys-Gln, Glu-Val, His-Phe, Ile-Tyr, Lys-Tyr, Lys-Val, Pro-Trp
14	5.082	4.731	0.791	18	3163	Ala-Ala, Ala-Arg, Ala-Leu, Ala-Tyr, Arg-Gln, Arg-Ile, Arg-Thr, Asp-Gln, Asp-Leu, Gln-Tyr, Glu-Glu, Glu-Ile, Glu-Met, Gly-Lys, His-Val, Ile-Phe, Ile-Val, Pro-Thr
15	5.095	4.898	0.780	7	977	Ala-Gln, Ala-His, Arg-Glu, Asn-Asp, Asp-Val, Cys-Tyr, Phe-Tyr
16	5.108	5.028	0.787	9	1503	Ala-Lys, Ala-Trp, Asp-Tyr, Gln-Glu, Glu-Ser, Gly-Met, Ile-Leu, Leu-Val, Met-Val
17	5.120	5.435	0.723	8	1040	Arg-Asp, Asn-Gln, Asp-Phe, Cys-His, Gln-Ser, Glu-Tyr, Leu-Ser, Met-Trp
18	5.131	4.965	0.991	7	1082	Ala-Glu, Arg-Met, Cys-Leu, Gln-His, Glu-Lys, Leu-Leu, Lys-Ser
19	5.146	5.031	0.789	16	2625	Ala-Asn, Ala-Asp, Ala-Met, Ala-Pro, Arg-Tyr, Asn-His, Asp-Met, Cys-Glu, Gln-Leu, Glu-Leu, Glu-Pro, Glu-Thr, Ile-Met, Ile-Pro, Leu-Lys, Leu-Tyr
20	5.161	5.687	0.548	12	1925	Ala-Gly, Asn-Cys, Asn-Glu, Asn-Tyr, Asn-Val, Asp-His, Gly-Leu, His-Leu, Lys-Phe, Met-Phe, Met-Ser, Thr-Val
21	5.172	6.123	0.678	7	1211	Asn-Leu, Asp-Thr, Gln-Lys, His-Lys, Ile-Ser, Leu-Thr, Lys-Met
22	5.192	6.057	0.604	12	1447	Ala-Phe, Ala-Ser, Arg-Trp, Cys-Cys, Glu-His, His-Thr, Ile-Trp, Lys-Thr, Met-Pro, Pro-Val, Ser-Thr, Tyr-Val
23	5.209	5.907	0.704	7	1017	Arg-Ser, Asn-Gly, Asn-Lys, Asp-Lys, Gln-Phe, Gln-Val, His-Ser
24	5.226	6.010	0.635	8	852	Asn-Trp, Asp-Asp, Asp-Cys, Asp-Ser, Cys-Ile, Glu-Trp, Ile-Thr, Ser-Ser
25	5.247	5.428	0.498	9	1260	Asn-Phe, Asp-Trp, Cys-Phe, Glu-Gly, Gly-Val, His-His, His-Ile, Leu-Pro, Ser-Val
26	5.270	5.957	0.511	2	391	Asp-Gly, Cys-Ser
27	5.291	6.321	0.480	7	747	Asn-Met, Asn-Thr, Cys-Lys, Gln-Gly, Gly-Tyr, Met-Tyr, Thr-Thr
28	5.342	6.010	0.437	5	627	Asn-Ser, Gly-Gly, His-Met, Lys-Pro, Pro-Tyr
29	5.369	6.382	0.437	9	1548	Arg-Gly, Arg-Pro, Cys-Val, Gly-Pro, Gly-Ser, Gly-Thr, Gly-Trp, Phe-Ser, Ser-Trp
30	5.427	6.861	0.403	4	265	Cys-Thr, Cys-Trp, Gly-His, Phe-Pro
31	5.493	5.883	0.491	1	63	His-Pro
32	5.651	5.823	0.164	1	68	Cys-Pro
33	5.727	6.873	0.733	1	82	Asn-Asn
34	5.899	6.639	0.164	1	127	Cys-Gly
35	5.985	6.819	0.414	1	48	Pro-Pro
36	8.001	8.669	0.854	1	8	Trp-Trp

[a] Energy in arbitrary units.

[b] Number of interaction types forming a class.

[c] Number of times the interactions occur in the calibration set of proteins.

TABLE VII. Long-range Interaction Parameters

	σ^{long} (Å)		ϵ^{long}			
Class	Initial	Final	(energy)[a]	N_{mem}[b]	N_{occur}[c]	Members
1	5.128	5.081	0.053	2	53 846	Cys-Cys, Gly-Gly
2	5.719	5.225	0.054	21	801 072	Ala-Gly, Ala-Ser, Arg-Gly, Asn-Gly, Asp-Gly, Cys-Gly, Cys-Phe, Gln-Gly, Gly-Phe, Gly-Pro, Gly-Ser, Gly-Thr, Gly-Trp, Gly-Tyr, His-Trp, Pro-Pro, Pro-Ser, Pro-Trp, Thr-Thr, Trp-Trp, Trp-Tyr
3	5.958	5.500	0.056	31	982 810	Ala-Met, Arg-Thr, Asn-Lys, Asn-Pro, Asp-Lys, Cys-Met, Cys-Thr, Cys-Trp, Cys-Tyr, Gln-Ser, Gln-Thr, Glu-Lys, Gly-His, Gly-Ile, Gly-Leu, Gly-Lys, Gly-Met, Gly-Val, His-His, His-Lys, Lys-Lys, Lys-Pro, Lys-Thr, Lys-Trp, Lys-Tyr, Met-Pro, Phe-Phe, Ser-Ser, Ser-Thr, Ser-Trp, Ser-Tyr
4	6.093	6.031	0.057	28	910 175	Ala-Cys, Ala-Ile, Ala-Pro, Arg-Asp, Arg-Cys, Arg-Pro, Arg-Ser, Arg-Tyr, Asn-Asn, Asp-His, Asp-Phe, Asp-Thr, Cys-Ile, Cys-Pro, Gln-Gln, Glu-Gly, His-Phe, Ile-Met, Ile-Tyr, Ile-Val, Lys-Ser, Phe-Thr, Phe-Val, Ser-Val, Thr-Trp, Thr-Tyr, Thr-Val, Val-Val
5	6.176	6.122	0.058	22	739 394	Ala-Ala, Ala-Asp, Ala-Thr, Ala-Tyr, Arg-Arg, Arg-Asn, Arg-Glu, Arg-His, Arg-Ile, Arg-Trp, Asn-Phe, Asn-Tyr, Glu-Ser, Glu-Thr, Ile-Ile, Met-Phe, Phe-Pro, Phe-Ser, Pro-Thr, Pro-Tyr, Tyr-Tyr, Tyr-Val
6	6.257	6.473	0.059	27	889 141	Ala-Glu, Ala-Leu, Ala-Phe, Ala-Trp, Ala-Val, Asn-Glu, Asn-Ser, Asp-Met, Asp-Ser, Cys-Leu, Cys-Val, Gln-His, Gln-Phe, Gln-Pro, Glu-Glu, His-Met, His-Thr, His-Tyr, Ile-Leu, Ile-Trp, Leu-Met, Leu-Trp, Leu-Tyr, Leu-Val, Phe-Trp, Pro-Val, Trp-Val
7	6.344	7.362	0.060	28	933 553	Ala-Arg, Ala-Gln, Ala-Lys, Arg-Val, Asn-Asp, Asn-Cys, Asn-Thr, Asp-Pro, Cys-Ser, Gln-Trp, Gln-Val, Glu-Ile, Glu-Met, Glu-Phe, Glu-Tyr, His-Pro, His-Ser, Ile-Lys, Ile-Phe, Ile-Ser, Ile-Thr, Leu-Phe, Leu-Ser, Lys-Phe, Met-Ser, Met-Thr, Met-Tyr, Met-Val
8	6.470	7.232	0.060	30	831 759	Ala-Asn, Arg-Gln, Arg-Leu, Arg-Met, Arg-Phe, Asn-Gln, Asn-His, Asn-Ile, Asn-Met, Asn-Trp, Asn-Val, Asp-Asp, Asp-Gln, Asp-Trp, Cys-Gln, Cys-Lys, Gln-Glu, Gln-Ile, Gln-Lys, Gln-Met, Gln-Tyr, Glu-His, Glu-Val, His-Ile, Ile-Pro, Leu-Leu, Leu-Pro, Lys-Val, Met-Trp, Phe-Tyr
9	6.662	7.833	0.061	19	665 929	Ala-His, Arg-Lys, Asn-Leu, Asp-Glu, Asp-Ile, Asp-Leu, Asp-Tyr, Asp-Val, Cys-Glu, Cys-His, Gln-Leu, Glu-Pro, Glu-Trp, His-Leu, His-Val, Leu-Lys, Leu-Thr, Lys-Met, Met-Met
10	6.942	7.657	0.063	2	81 302	Asp-Cys, Glu-Leu

[a] Energy in arbitrary units.

[b] Number of interaction types forming a class.

[c] Number of times the interactions occur in the calibration set of proteins.

structures. The worse results are obtained for 1ctx, a 71 residue, snake toxin whose conformation is dominated by five disulfide bridges, rather than regular secondary structure. The other weak results are for 1pcy, the apo form of plastocyanin, which normally has an integral copper atom in situ. It is also clear that the nearest local minima really are near to the native structure. The last column of Table VIII shows that no structure moves by more than 0.2 Å (distance matrix difference) upon minimizing. To the extent that 10 proteins represent a significant test, the force field really is reflecting general properties of protein folds, since all the test proteins were not homologous to proteins used in the parametrization.

TABLE VIII. Testing of Force Field Generalization

			Low-energy alternatives				
			Before minimization		After minimization		rms shift [e]
Protein	N_{res}[a]	$N_{\mathrm{alternatives}}$[b]	N[c]	%[d]	N[c]	%[d]	(Å)
4pti	58	37 878	57	0.2	1	0.0	0.07
1sn3	65	36 024	3 172	8.8	1	0.0	0.13
1ctx	71	34 501	11 699	33.9	1 489	4.3	0.15
1pcy	99	27 998	3 137	11.2	498	1.8	0.12
1lyz	129	22 558	1 275	5.7	0	0.0	0.16
4fxn	138	21 130	8 967	42.4	0	0.0	0.13
2sns	141	20 662	14 062	68.1	25	0.0	0.14
1rhd	293	5 286	5 249	99.3	0	0.0	0.15
1abp	306	4 571	4 565	99.9	0	0.0	0.17
5cpa	307	4 516	0	0.0	0	0.0	0.18
3tln	316	4 047	43	1.1	0	0.0	0.17

[a] Number of residues.

[b] Number of alternative conformations generated.

[c] Number of alternatives with energy lower than that of the native structure.

[d] Percentage of alternatives with energy lower than that of the native structure.

[e] Root-mean-square distance matrix difference of the native structure before and after minimization.

These results are disappointing in that they suggest that a force field of this type would be slow for practical threading applications as it would only be useful after energy minimization. It also suggests that the inverse twelfth power repulsive terms in Equations (2)–(4) increase too steeply. In some cases, the native structure had a huge E^{long} contribution mostly due to just one or two pairs of residues at a short distance (data not shown). This result is not a surprise, given this force field, but it is not clear why similar results are not seen with table-driven force fields[2,31] based on a simple Boltzmann relation or perhaps whether they do sometimes occur. Amino acid pairs are sometimes present at unusually short distances in native structures. Statistically, these should be poorly represented and should give rise to correspondingly high energies in table-based force fields.

DISCUSSION

The results show the feasibility of an unusual method for force field parametrization, although this implementation was certainly not ideal. Little attempt was made to optimize technical details such as the dynamics temperature or temperature coupling. Unfortunately, it is almost impossible to know ideal values for constants that define the dynamic behavior in parameter space. One can compare this with the field of protein molecular dynamics (MD) simulations that is relatively mature, but where analogous quantities are often chosen from experience rather than rational forethought. A simulation in parameter space should be based on the shape of the parameter energy hypersurface, but, to continue the comparison with protein simulations, this surface is still not well understood in the field of protein MD.[29] The weakest point of this calculation is probably not whether or not the some control value was optimal, but rather the short length of the calculation; 100 time steps is not even a cursory peek in parameter space, let alone the thorough search one would like. It is certainly not enough to see convergence of the parameters. One strength of the optimization procedure here was the selection of alternative structures for calculating forces. From one point of view, this is an approximation to make the calculations faster; only alternative structures of low energy contributed to the force calculation. At the same time, there is a more subtle benefit. Every alternative structure contributes local minima and maxima to the total parameter energy surface. Removing less important alternative structures must smooth the energy surface and make searching easier. Obviously, this is an observation of principle rather than a quantitative statement.

There are also some aspects of the parameter optimization that have a systematic effect on parameters. For example, a scheme without any scaling of contributions from different proteins would be influenced most by the native structures of larger proteins (which contain the most interactions) and the alternative structures of smaller proteins (which have have the most misfolded alternatives). The scheme used here has replaced this by a different bias [Eq. (11)].

The parametrization scheme and use of misfolded structures may also limit the application areas of the force field. By definition, a conventional molecular mechanics force field has its minima located at positions of minimum energy. This force field differs in that parameters are not chosen solely on the basis of native structures. Minima are, of course, located at pseudoenergy minima, but these are positioned so as to optimize discrimination ability, rather than reproduce native structures. This means that mathematically, one could use the force field for newtonian simulations, but the structures of low pseudoenergy may not be physically realistic configurations. This also means that the force field may not be useful for looking for errors in structures, an application area possible with other protein scoring functions.[5,30] This type of force field may be useful for grossly wrong structures such as mistraced crystallographic density, since that kind of error will produce compact misfolded structures similar to the parametrization data. It may not be so useful detecting smaller errors in structures.

The effect of other implementation choices is less clear. During the parameter dynamics, forces were calculated based on crystallographic coordinates of native structures. As shown in the Results section, energies of structures are only useful after energy minimizing. This is clearly a disadvantage when compared to table-based force fields, which produce good results directly from crystallographic coordinates.[2,31]

While one can speculate on the effects of the optimization method, a more fundamental question is the suitability of the form of interaction functions. Clearly, the set of functional forms has limitations, but it is not clear what the weakest aspects are and where it could be most profitably enhanced. As described in the Results section, the twelfth power repulsive term increases too quickly. Some weaknesses are highlighted by other workers. For example, our force field uses isotropic interactions between C^α atoms, but this is only a poor representation of the anisotropic interactions of real amino acids.[31] The single-well interaction functions [Eqs. (2)– (4)] are not an ideal fit to the data when one considers published distributions of amino acid pair distributions,[2,32] and the exponents in the potential energy terms are also quite arbitrary. One could well argue for 8–6 exponents[14,25] or 12–10 exponents.[19] There are even more fundamental questions about the most appropriate formulation. For example, we have an n, $n + 3$ interaction where the choice of parameters depends on the central ($n + 1$ and $n + 2$) amino acids.[19,25] This recognizes that amino acids have a statistical preference for the middle of specific secondary structures. By contrast, other workers have chosen parameters for the n, $n + 3$ interaction based on the outer (n and $n + 3$) residues.[2,31] It might appear that this is an irreconcilable difference, but the best answer probably lies between the approaches. Considering the example of the n, $n + 3$ term, there must be a contribution from both central and outer residues to the interaction. Given a tractable methodology for generating force fields, we hope to resolve points like this in a quantitative manner.

Possibly, the biggest omission is the lack of an explicit solvation potential energy term. In fact, there is evidence that this kind of term alone may be successful in recognizing correct folds.[2,33,34] Solvation effects are included in our force field in the same way as all structural influences affect the final parameters. The weakness is that our Lennard-Jones-like interactions may not be well suited to include the influence of solvent. If one wanted to justify our nonoptimal functional form, it is interesting to note that an explicit solvent term based on a property such as solvent-exposed surface area may not be necessary, and some workers have achieved remarkable results modeling the hydrophobic effect with very few parameters.[35,36,37]

Viewing the force field generation as an exercise in model fitting, the effect of a badly chosen functional form is clear. The fitting has been conducted on a small set of proteins and reproduces the data well within that range. The worse the functional form, the less likely the fitting is to be useful outside of that range.

The functional form also places limitations on the areas of application, as does the parametrization methodology, discussed above. With interaction sites at only C^α positions, this is a low-resolution force field and not necessarily sensitive to small changes or errors in structure. The force field is also achiral, so it will never be able to distinguish between mirror images such as right- or left-handed helices. This is obviously not a problem if it is only challenged with alternate structures generated from other native folds.

The force field parameters are also a reflection of the choice of proteins in the calibration set and the experimental conditions used in structure determination. This means that the results are influenced by a range of pH, salt concentrations, and so on. Nevertheless, the simple force field is rather adept at recognizing correct structures, regardless of their origin or experimental conditions. Presumably, the force field is biased toward soluble globular proteins under conditions most often used by crystallographers. It may well produce wrong answers outside of this set.

The success of a simple force field is the more surprising, considering the physical factors that go toward protein folding. If one could state that all the protein structures were at free-energy minima, then one could say that the force field is a fitting to underlying physical principles. If one believes that some protein structures are kinetically trapped local minima, then the force field is an ill-defined mixture of energetic principles and the kinetics of protein folding.

One can continue reasoning in this vein and note other influences that are neglected in the force field. We assume there are 20 amino acids interacting without any interference, but this is probably not the case. The set of proteins probably includes some conformations that are only viable due to bound metal ions, and there may be sets of acidic residues that are unusually close to each other due to coordinated anions. Many proteins include poorly defined residues whose coordinates are simply a reflection of the crystallographic refinement programme. Finally, one should note that, even though we applied criteria based on resolution and R factors, a set of about 100 structures almost certainly includes misinterpreted electron density and maybe even mistakes in amino acid sequences.

CONCLUSIONS

This work has not produced a final force field for protein structure recognition. The tests on proteins outside of the calibration set show too many false positives (low energy structures) to be of direct practical use. Energy minimization is required to obtain significant discrimination between native and misfolded structures. The work does, however, contain a number of improvements over other approaches to determining force fields. The use of quasi-newtonian dynamics is interesting because there is an implicit assumption that there is a volume of parameter space that will provide an acceptable force field and that one only needs to find some point within that volume. The overall form of the force field is not innovative, but some aspects are a distinct improvement over other approaches. Classifying interaction types rather than residues allows more flexibility in the fitting operation without increasing the number of parameters and the parameter classification algorithm is a rational method for decreasing the number of adjustable parameters.

A truly systematic exploration of force field functional forms will probably be beyond the scope of practical computations for some time. However, the framework developed here should allow a reliable and reproducible scheme for building and testing simple force fields for protein structure prediction.

REFERENCES

1. Jones, D., Thornton, J. Protein fold recognition. J. Comp-Aided Mol. Design 7:439–456, 1993.
2. Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding: An approach to the computational determination of protein structures. J. Comp.-Aided Mol. Design 7:473–501, 1993.
3. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3:247–259, 1993.
4. Bryant, S.H., Altschul, S.F. Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. 5:236–244, 1995.
5. Sippl, M.J. Recognition of Errors in three-dimensional structures of proteins. Proteins 17:355–362, 1993.
6. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. J. Mol. Biol. 232:805–825, 1993.
7. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
8. Covell, D.G., Jernigan, R.L. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.
9. Skolnick, J., Kolinski, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.
10. Dill, K.A. Folding proteins: Finding a needle in a haystack. Curr. Opin. Struct. Biol. 3:99–103, 1993.
11. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352 (1994).
12. Godzik, A., Kolinski, A., Skolnick, J. Lattice representations of globular proteins: How good are they? J. Comput. Chem. 14:1194–1202, 1993.
13. Levitt, M., Warshel, A. Computer simulation of protein folding. Nature 253:694–698, 1975.
14. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104:59–107, 1976.
15. Tanaka, S., Scheraga, H.A. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9:945–950, 1976.
16. Kuntz, I.D., Crippen, G.M., Kollman, P.A., Kimelman, D. Calculation of protein tertiary structure. J. Mol. Biol. 106:983–994, 1976.
17. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
18. Berger, J.O. "Statistical Decision Theory and Bayesian Analysis." New York: Springer-Verlag, 1985.
19. Crippen, G.M., Snow, M.E. A 1.8 Å resolution potential function for protein folding. Biopolymers 29:1479–1489, 1990.
20. Maiorov, V.N., Crippen, G.M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
21. Ryckaert, J.P., Ciccotti, G., Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. J. Comput. Phys. 23:327–341, 1977.
22. Allen, M.P., Tildesley, D.J. "Computer Simulation of Liquids." Oxford: Clarendon Press, 1987.
23. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Prot. Sci. 3:522–524, 1994.
24. Crippen, G.M. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 30:4232–4237, 1991.
25. Seetharamulu, P., Crippen, G.M. A potential function for protein folding. J. Math. Chem. 6:91–110, 1991.
26. Havel, T.F. The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: A study of 1830 independently computed conformations. Biopolymers 29:1565–1585, 1990.
27. Maiorov, V.N., Crippen, G.M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J. Mol. Biol. 235:625–634, 1994.
28. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., Haak, J.R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81:3684–3670, 1984.
29. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.
30. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356:83–85, 1992.
31. Kocher, J.-P.A., Rooman, M.J., Wodak, S.J. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol. 235:1598–1613.
32. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. Proteins 6:193–209, 1989.
33. Abagyan, R., Frishman, D., Argos, P. Recognition of distantly related proteins through energy calculations. Proteins 19:132–140, 1994.
34. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 7:257–264, 1990.
35. Sun, S., Thomas, P.D., Dill, K.A. A simple protein folding algorithm using a binary code and secondary structure constraints. Prot. Eng. 8:769–778, 1995.
36. Srinivasan, R., Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.
37. Huang, E.S., Subbiah, S., Levitt, M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J. Mol. Biol. 252:709–720, 1995.