

Potential Energy Function for Continuous State Models of Globular Proteins

Y. ZENMEI OHKUBO and GORDON M. CRIPPEN

ABSTRACT

One of the approaches to protein structure prediction is to obtain energy functions which can recognize the native conformation of a given sequence among a zoo of conformations. The discriminations can be done by assigning the lowest energy to the native conformation, with the guarantee that the native is in the zoo. Well-adjusted functions, then, can be used in the search for other (near-) natives. Here the aim is the discrimination at relatively high resolution (RMSD difference between the native and the closest nonnative is around 1 Å) by pairwise energy potentials. The potential is trained using the experimentally determined native conformation of only one protein, instead of the usual large survey over many proteins. The novel feature is that the native structure is compared to a vastly wider and more challenging array of nonnative structures found not only by the usual threading procedure, but by wide-ranging local minimization of the potential. Because of this extremely demanding search, the native is very close to the apparent global minimum of the potential function. The global minimum property holds up for one other protein having 60% sequence identity, but its performance on completely dissimilar proteins is of course much weaker.

Key words: protein folding, potential energy function, threading, Boltzmann statistics, global minimum.

INTRODUCTION

THERE HAVE BEEN MANY COMPUTATIONAL STUDIES related to protein folding that are concerned primarily with the connection between amino acid sequence and three-dimensional structure, rather than folding kinetics, folding pathways, etc. In these studies the central feature is the potential function that is supposed to represent the free energy of the protein/solvent system as a function of protein sequence and conformation. Except for all-atom, explicit solvent molecular dynamics calculations, the potential is the free energy of the system integrated over solvent degrees of freedom and minor vibrational degrees of freedom. Of course the potential must also be compatible with the way the polypeptide chain is represented in the calculations, which may range from isotropically interacting point residues on a square lattice to polarizable point atoms in continuous three-dimensional Euclidean space.

One way to study the relationship between protein sequence and structure is to begin with some hypothesized potential function suitable to the polypeptide chain representation and a very restricted, often discrete conformation space. Then the sequences that fold up are those having a unique global minimum conformation for the given potential function, and that conformation corresponds to the native structure. If

the conformation space is small enough so that these global optimizations of the potential are feasible, one can study the consequences of the model (e.g., Yue and Dill, 1995; Deutsch and Kurosky, 1996; Crippen and Ohkubo, 1998) and try to extend the conclusions to the behavior of real proteins.

The opposite approach is to use experimental evidence about real proteins, primarily their crystal structures, to construct a potential function that agrees with this evidence in some sense. These studies differ in the representation of the polypeptide chain, the functional form of the potential, the algorithm for adjusting the parameters of the potential, the set of sequences under consideration, the types of conformations examined, and the measures of success employed. It is not surprising there is considerable disagreement and confusion in the field. Ultimately one application of these potentials would be to predict the native structure of a given sequence. Here we focus on a restricted version of this problem, variously called "3DID," "sequence seeks structure," or "fold recognition." Namely, given a sequence, the potential function, and a set of structures that includes the native, the potential should give the lowest/best value for the native structure, compared to all the other nonnative structures. After all, this is what a real protein does when it reversibly folds up under thermodynamic control.

Many ways have been proposed to devise potentials for fold recognition. There are comparisons of those methods (e.g., Godzik *et al.*, 1995; Wang *et al.*, 1995; Park and Levitt, 1996; Park *et al.*, 1997), and reviews (e.g., Torda, 1997; Vajda *et al.*, 1997; and references therein). Consider the conceptual classification illustrated schematically in Figure 1, where sequence and conformation are multidimensional, large, but finite axes, and the potential function axis is mathematically infinitely dimensional. Some potentials arise from general principles observed in protein structures and few adjustable parameters (Huang *et al.*, 1995). A strictly knowledge-based potential of mean force can be derived from a broad survey of the crystal structures of many different proteins (Miyazawa and Jernigan, 1985, 1996; Hendlich *et al.*, 1990; Nishikawa and Matsuo, 1993), as indicated by the bar marked (a). It has been argued that such a method produces artifacts due to the population of structures surveyed, and it does not reproduce the correct potential in artificial test cases where the potential is known by construction (Thomas and Dill, 1996a). However, a comprehensive comparison of native and nonnative folds for many folding sequences can reproduce the correct potential, either by an iterative method (Thomas and Dill, 1996b) or by solving a large set of inequalities (Crippen, 1996; Maiorov and Crippen, 1992). The space of possible potential functions is large, and even in these artificial test cases that examine all possible sequences and structures, the potential used in setting up the problem is only one of a diverse set of such functions that can always favor the native conformation over all nonnatives for every sequence that folds to a unique native structure (Crippen, 1996).

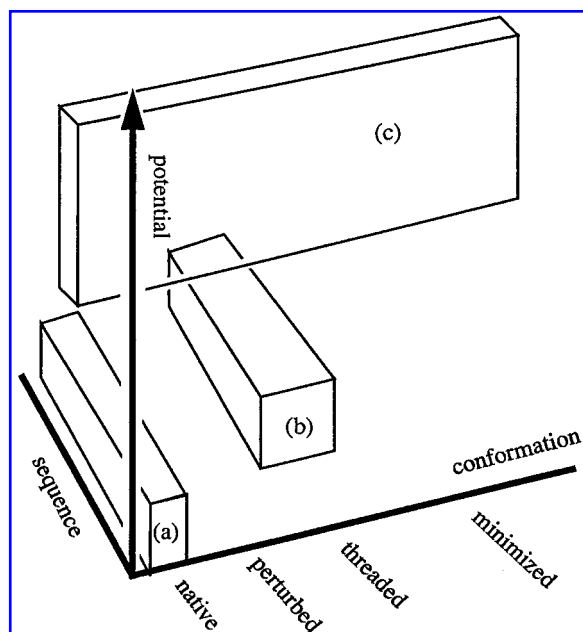


FIG. 1. A conceptual view of different potentials for fold recognition in terms of the sequences and structures used for training, and the sorts of potential functions employed.

A broadly applicable potential function derived by comparing real protein native and nonnative structures (Maiorov and Crippen, 1992) relied on generating the nonnatives by threading (Hendlich *et al.*, 1990), where the ungapped native sequence is applied to a contiguous piece of a larger protein structure. This is represented in Figure 1 by bar (b) in that it used different structures for training, the functional form of the potential was different, and the set of native sequences for training was broad but necessarily had to exclude some proteins that were stabilized by effects outside the scope of the model, such as lipid membranes or essential prosthetic groups. In that work, the parameters of the potential were adjusted so as to satisfy a set of linear inequalities, namely, that the value of the potential for the native structure should be lower than that for the (threaded) nonnative. Other ways of building in fold recognition of native structures compared to some population of nonnatives include maximizing the Boltzmann probability of the native conformations (Hao and Scheraga, 1996), maximizing the negative of the harmonic mean of Z scores (Mimiy and Shakhnovich, 1996), maximizing the statistical weight of the natives (Thomas and Dill, 1996b), or using linear programming to minimize the worst violation of the recognition inequalities (Akutsu and Tashimo, 1998). Z scores (Bowie *et al.*, 1991) express how many standard deviations the energy of natives is from the average energy of the nonnatives.

Threading alone, however, may be too loose a condition for training (or checking). Since the structural templates are all good protein crystal structures, the potential is never asked to differentiate between good packing vs. steric overlaps, or between left- and right-handed helices. The natives may not have lower energies than all nonnatives, because the number of conformations to be checked for each native is limited. Besides, threaded conformations may have relatively high energies, because those conformations are expected to be stable for their own sequence (and ones with minor mutations) and not necessarily for the threaded sequence. Potentials trained by threading might give lower energies to nonnatives around the native conformation ("near-natives") than the native itself. Levitt and coworkers pointed out these shortcomings and took a challenging approach; they generated nonnatives around the native by molecular dynamics (perturbed conformations in Figure 1) and checked whether the native conformation had a lower energy than the low-energy nonnatives by their potential (Huang *et al.*, 1996). This condition is much more stringent than threading training, as nonnatives by threading are at relatively high energies (see Figure 1 in Vendruscolo and Domany [1998]). Levitt and coworkers found their potentials worked well but not perfectly, and consequently they proved that training by threading only is a rather loose necessary condition for the recognition problem.

In the same way, it is quite natural to suspect that the threaded conformations may be located on the slopes of the energy surface and that there are local minima nearby which have a lower energy than the native. Besides, the native itself may not even be at a local minimum. Therefore, the approach we take here is: (1) train our potential by threading, (2) generate energy-minimized nonnatives by local energy minimizations starting from the native and randomly chosen threaded conformations, and then (3) adjust the parameters of our potential so that the native is at a local minimum and has a lower energy than the low-energy nonnatives. In other words, the native is apparently at the global minimum. Steps 1–3 are repeated until eventually we find no conformation which has a lower energy than the native. Clearly this is an extremely laborious process due to the numerous energy minimizations, and at the outset it was not clear that even our relatively flexible functional form would succeed in always recognizing the native compared to such exceptionally good nonnatives. Consequently, in this work we concentrated on a wide range of potential functions and a single native sequence vs. the full panoply of types of nonnative conformations, symbolized by bar (c) in Figure 1.

This procedure is similar to that of Crippen (1996). Only three things are needed: sequences and conformations, conformational similarity metric, and potential functions which are linear in the adjustable parameters. Here we are trying to derive a relatively simple potential energy function which depends on interaction distance and atom types in a continuous internal coordinate space under the conditions of fixed bond lengths and angles. The potentials are not based on the quasichemical approximation (Miyazawa and Jernigan, 1985, 1996), empirical knowledge, or any specific assumption, such as setting nonnative interactions at neutral in the Gō model (Gō, 1983). The only condition we require is that the function should always give the lowest value to the native conformation so as to find it among a zoo of conformations.

We employ pairwise, continuous, and additive functions and then adjust the parameters by the procedure outlined above. This functional form is based on the assumption that united-atom-level interaction energy can account for solvation effects. Also, the functions are independent of temperature, assuming normal laboratory temperatures. Since the conformations we employed are in terms of ϕ s and ψ s, it is not convenient to treat disulfide bridges, ligands or multiple polypeptide chains, and we did not include the

interaction among those. We expect, therefore, the resulting energy function should work well for globular monomeric proteins without ligands or disulfide bonds at room temperature only. We do not expect the function will work well for membrane proteins. Because the energy function favors the native over the nonnatives, one can regard it as an approximation to the free energy, or one can regard such a function as just a scoring function to find the native conformation.

METHODS

Continuous state model

We selected 313 X-ray-determined, monomeric proteins of no longer than 250 residues without big chain breaks or ligands (Table 1) out of the 25% list of PDB_Select (December 1998 release, Hobohm and Sander, 1994). To reduce the size of the conformation space, we fitted each of the PDB structures (Bernstein *et al.*, 1977; Berman *et al.*, 2000) to a standard geometry polypeptide model having all *trans* peptide bonds. The fitted model consists of main chain heavy atoms and C_{β} s (pseudo- C_{β} for Gly) with the standard values of bond lengths and angles (Ramachandran *et al.*, 1974); the (ϕ, ψ) values are any real numbers ranging from -180.0° to 180.0° . This model is a simplified version of that for ECEPP (Momany *et al.*, 1975), or a continuous version of the model of Park and Levitt (1995), and is substantially (there may be minor differences in fixed bond angles and lengths used) identical to the one used by Dill and coworkers (1997). The average RMSD (Kabsch, 1978) between a PDB structure and the fitted model is less than 0.5 \AA .

TABLE 1. 313 PDB ENTRIES CHOSEN FROM THE 25% LIST OF THE DECEMBER 1998 RELEASE OF PDB_SELECT (HOBOHM AND SANDER, 1994)

119l	153l	1a17	1a1x	1a68	1a6g	1a7i	1aa0	1ab7	1aba
1acp	1acz	1ad2	1ad6	1afp	1ag4	1agg	1ah1	1ah7	1ah9
1ahk	1aho	1aie	1ail	1aj3	1ajj	1akz	1al3	1aly	1amm
1amx	1an8	1aol	1aoo	1aoy	1ap0	1ap8	1apf	1apj	1aqb
1ark	1ash	1asx	1atg	1awd	1awj	1awo	1ax3	1b10	1bak
1bam	1baq	1bc4	1bcn	1bct	1bd8	1bdo	1be1	1bea	1bei
1beo	1bf8	1bfg	1bgf	1bkf	1ble	1bol	1bor	1br0	1brf
1bsn	1btn	1buz	1bv1	1bvh	1bw3	1bxa	1bym	1c25	1c52
1c5a	1cby	1cdb	1cdi	1cex	1cfb	1cfe	1cfh	1chd	1chl
1cid	1ctj	1cto	1cur	1eyo	1cyx	1dad	1ddf	1dec	1def
1dfx	1dhr	1div	1dun	1eal	1eca	1ehs	1erd	1erv	1exg
1fbr	1fna	1fua	1fus	1gky	1gps	1grx	1gvp	1hcd	1hev
1hfc	1hfh	1h1b	1hoe	1hqi	1ido	1ife	1ife	1irl	1jer
1jli	1jpc	1juk	1jvr	1kbs	1kid	1knb	1kpf	1krt	1ksr
1kte	1kuh	11ba	11bu	11cl	11eb	11it	11ki	11ou	11rv
1mai	1mak	1mb1	1mbh	1mbj	1mrj	1msc	1msi	1mup	1mut
1mzm	1ngr	1nkl	1nkr	1nls	1noe	1nox	1npk	1nxb	1ocp
1ois	1opd	1opr	1orc	1pce	1pdo	1pex	1pft	1pih	1pkp
1plc	1pne	1poa	1poc	1pou	1ppn	1ppt	1put	1qyp	1ra9
1ref	1ret	1rie	1rlw	1rmd	1rof	1rpo	1rsy	1sco	1sfe
1sfp	1skz	1spy	1sra	1sro	1std	1svr	1tam	1tbn	1tfb
1tfe	1thv	1tih	1tit	1tiv	1tle	1tpn	1tsg	1tul	1ubi
1ulo	1utg	1uxd	1vcc	1vhh	1vid	1vif	1vig	1vls	1vsd
1vtx	1wab	1whi	1who	1wui	1wkt	1xnb	1ycc	1yua	1yub
1zaq	1zin	1zug	1zwa	1zxq	2a0b	2abd	2abk	2acy	2adx
2ayh	2baa	2bb8	2bby	2bds	2brz	2cps	2ech	2end	2eng
2erl	2ezh	2ezl	2fdn	2fha	2fn2	2fow	2fsp	2gdm	2hbg
2hfh	2hgf	2hoa	2hp8	2hqi	2i1b	2igd	2ilk	2lbd	2lfb
2mcm	2nef	2new	2pac	2phy	2pii	2pth	2ptl	2pvb	2rgf
2rn2	2sak	2sn3	2sns	2stv	2sxl	2tbd	2tgi	2ucz	2vgh
2vil	3bbg	3chy	3cla	3cyr	3lzt	3nll	3seb	3vub	4mt2
5p21	5pti	7rsa							

Fitting a PDB structure to a continuous-state model

The fitting of PDB structures to the standard-geometry continuous-state model is carried out by minimizing a penalty function, P :

$$\begin{aligned}
 P = & \sum_{length} (d_{ij}^2 - d_0^2)^2 \\
 & + \sum_{angle} (\theta_{ijk} - \theta_0)^2 \\
 & + \sum_{\omega} (\omega_i - \pi)^2 + \sum_{\omega} (coplanar)_i^2 \\
 & + \sum_{atom} (c_i - c_{i,PDB})^2
 \end{aligned} \tag{1}$$

where i, j, k signify any main chain heavy atoms or C_{β} of the current conformation. Therefore, d_{ij} is the bond length between any bonded atom pair i and j ; d_0 is the standard bond length of the i and j pair (Schulz and Schirmer, 1979); θ_{ijk} is the bond angle of any bonded atom triple $i-j-k$; θ_0 is the standard bond angle of the corresponding atoms; ω_i is the i -th peptide bond dihedral angle of the conformation; π is set at the standard dihedral angle of ω ($trans = 180^\circ$); c_i is the coordinate of atom i ; and $c_{i,PDB}$ is the coordinate of the atom i in the PDB structure. The $(coplanar)_i$ term is for the i -th ω dihedral angle:

$$(coplanar)_i = \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_{C_i^\alpha} & x_{C_i'} & x_{N_{i+1}} & x_{C_{i+1}^\alpha} \\ y_{C_i^\alpha} & y_{C_i'} & y_{N_{i+1}} & y_{C_{i+1}^\alpha} \\ z_{C_i^\alpha} & z_{C_i'} & z_{N_{i+1}} & z_{C_{i+1}^\alpha} \end{vmatrix} \tag{2}$$

where $x_{C_i^\alpha}$, for instance, signifies the x coordinate of the i -th C_{α} , and so on. The squared coplanar term is 0 when all the four atoms lie in the same plane and increases steeply if any atom among the four deviates from the plane. The third term forces a *cis* configuration to convert to *trans*, and then the fourth term holds the ω dihedral angle at exactly 180° . Any *cis*-Pro, therefore, is converted to *trans*-Pro.

The conjugate gradient method (Hestenes, 1980, and references therein) is used to minimize P , starting from the PDB conformation. The minimization consists of two procedures: after 1,000 steps of minimization, the current conformation has a near-standard geometry and near-PDB conformation. Then another 5,000 step minimization is carried out without the last term of P , which allows convergence to an almost standard geometry conformation. The obtained (ϕ, ψ) values are used to build the standard geometry conformation. The rebuilt standard geometry conformation and the P -minimized one are substantially identical (RMSD between the two is around 0.01 Å or less), as P has been minimized to a pretty small value, namely, on the order of 10^{-5} or less. The average RMSD between a PDB structure and the fitted model is less than 0.5 Å.

Native and nonnative models

In the current work, the fitted model of ubiquitin (1ubi, 76 residues) is regarded as the native, and we used threaded conformations from all the other models (Table 1) plus 7,000 randomized ones from 1ubi as follows:

- 36,646 threaded conformations of 313 fitted models concatenated in alphabetical order of the PDB identifiers;
- 1,000 conformations of the fitted 1ubi whose (ϕ, ψ) pair at one randomly chosen residue is changed to an existing pair randomly chosen among 313 fitted models;
- same as in (b), except two (ϕ, ψ) pairs are randomized;
- same as in (b), except five pairs are randomized;
- same as in (b), except ten pairs are randomized;

- (f) same as in (b), except twenty pairs are randomized;
 (g) same as in (b), except all pairs are randomized;
 (h) same as in (g), except all pairs are randomly perturbed within a range of $\pm 10^\circ$.

The randomized nonnatives (b)–(d) tend to keep local native conformations, but do not hold global native topology. On the other hand, the perturbed nonnatives (h) tend to hold global native topology, but not local native conformations. The nonnatives (e) and (f) are intermediate. Nonnatives (a) and (g) keep neither global nor local native conformation. Also, nonnatives in those two categories tend to be totally dissimilar to the native, while (h) tend to be near-natives. The prepared nonnatives cover various kinds of conformations and a wide range of conformational similarity to the native. These nonnatives are used as starting points to get energy-minimized nonnatives (see *Quadratic Programming for Parameter Adjustment* Section).

Potential function

We employ a pairwise type potential energy function as an energy function which can select the native conformation of a given sequence out of an assortment of conformations. The functions are atom-type and atom-distance dependent, and each is expressed as a linear combination of chosen basis functions. The total energy of a conformation, E , is the sum of the energies of any 1–4 or further atom pair, whose distance will change when the dihedral angle is changed:

$$\begin{aligned}
 E &= \sum_{|i-j| \geq 4} e_{t_i t_j}(r_{ij}) \\
 &= \sum_{|i-j| \geq 4} \sum_{l=0}^3 x_{l t_i t_j} b_l(r_{ij})
 \end{aligned} \tag{3}$$

where $e_{t_i t_j}(r_{ij})$ is the interaction between atoms i and j at a distance of r_{ij} . t_i and t_j signify the atom types of i and j , respectively. We have tried various sets of basis functions, $b_l(r_{ij})$, and in this paper,

$$b_l = \begin{cases} ((r_{ij} - a)^2 - b^2)^2 / b^4 & \text{for } a - b \leq r_{ij} \leq a + b \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

is used with $a = 0.0, b = 4.0$ for b_0 ; $a = 4.0, b = 4.0$ for b_1 ; $a = 6.0, b = 4.0$ for b_2 ; $a = 8.0, b = 4.0$ for b_3 . $b_l(r_{ij})$ s reach 1.0 when $r_{ij} = a$, and 0.0 when $r_{ij} = a \pm b$ (Figure 2). The coefficients for $b_0, \{x_{0 t_i t_j}\}$, are fixed at 10.0, while the others, $\{x_{1 t_i t_j}\}, \{x_{2 t_i t_j}\}$, and $\{x_{3 t_i t_j}\}$, are to be adjusted between -10.0 and 10.0 , in order to obtain the native-discriminative energy function. That restricts each $e_{t_i t_j}(r_{ij})$ within a possible range of -21.25 to 21.25 and forces them to gradually reach zero at long distances, making the potential surface simple and smooth:

$$\begin{aligned}
 -21.25 &\leq e_{t_i t_j}(r_{ij}) \leq 21.25 \\
 e_{t_i t_j}(0.0) &= 10.0 \\
 \left. \frac{\partial}{\partial r_{ij}} e_{t_i t_j}(r_{ij}) \right|_{r_{ij}=12.0} &= e_{t_i t_j}(12.0) = 0.0
 \end{aligned} \tag{5}$$

Number of parameters

As described above, there are three adjustable parameters for each atom type pair: $x_{1 t_i t_j}, x_{2 t_i t_j}$, and $x_{3 t_i t_j}$. Since the number of atom types is nineteen (main chain heavy atoms and C_β s with their residue types, but V/L/I, S/T, D/E, and N/Q are grouped into the same type, respectively), there are 190 combinations of atom types. The total number of the adjustable parameters, therefore, is $3 \times 190 = 570$.

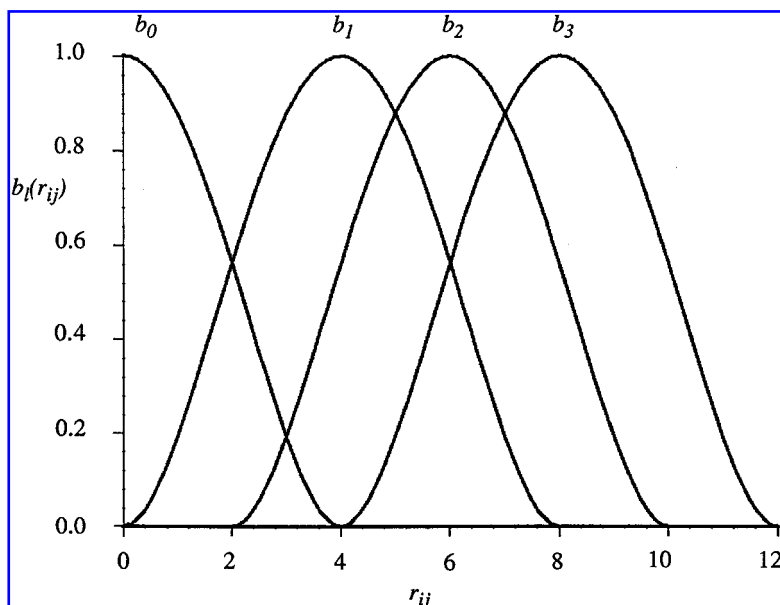


FIG. 2. The basis functions used: $b_0(r_{ij})$, $b_1(r_{ij})$, $b_2(r_{ij})$, and $b_3(r_{ij})$ from left to right.

Constraints

The requirement that each nonnative should have a higher energy than the native can be described as a set of inequalities:

$$\Delta E = E_{\text{nonnat}} - E_{\text{nat}} > g, \quad (6)$$

$$g = \begin{cases} 0.3 & \text{if } 0.3 \leq \rho \\ \rho & \text{if } 0.1 < \rho < 0.3 \end{cases}$$

where E_{nat} and E_{nonnat} signify the energy of the native and that of a nonnative, respectively. The quantity ρ is a size-independent metric of conformational similarity (Maiorov and Crippen, 1995) between the native and nonnative; ρ equals 0 if two conformations are identical, and ρ reaches 2 when two conformations are totally dissimilar (~ 1.8 in the case of a chain of equally spaced particles, such as these protein models). The quantity ρ is about one tenth of RMSD in case of 1ubi. The proportionality between g and ρ is set for smaller ρ so that the requirement is not too strict for “neighbor” nonnatives. Ordinarily, if ρ is less than 0.3, the nonnative keeps the native’s topology. As Vendruscolo and Domany (1998) point out, the conformations by threading tend to score a relatively high energy as they may not be located at the local minima of the energy surface, although threading itself serves as a good source of nonnative conformations. The inequality here, therefore, is generated not only for the nonnatives (a) through (h) but also for energy-minimized conformations using the current potential during training (see next section for details). Any conformation with ρ of less than 0.1 is excluded from generating an inequality, as the conformation is very close to the native and it can be regarded as one of the “natives.”

Quadratic programming for parameter adjustment

The adjustable parameters, $\{x_{1:t_j}\}$, are determined by quadratic programming (for reference, see Bazaraa *et al.* [1993]). Quadratic programming is a method to optimize (either minimize or maximize) a quadratic objective function subject to a set of linear equality and/or inequality constraints. A typical quadratic programming problem is:

$$\begin{aligned} &\text{Minimize} && \frac{1}{2} \mathbf{x}^t \mathbf{Q} \mathbf{x} + \mathbf{c}^t \mathbf{x} \\ &\text{subject to} && \mathbf{A} \mathbf{x} \geq \mathbf{b} \end{aligned} \quad (7)$$

where \mathbf{c} and \mathbf{x} are n dimensional column vectors, \mathbf{Q} is an $n \times n$ symmetric matrix, \mathbf{A} is an $m \times n$ matrix, and \mathbf{b} an m dimensional column vector. The domain $S = \{\mathbf{x} \in R^n \mid \mathbf{Ax} \geq \mathbf{b}\}$ is called the feasible region of this quadratic program. Quadratic programming finds the unique, optimal solution, the $\mathbf{x} \in S$ which gives the minimum value of $\frac{1}{2}\mathbf{x}^t\mathbf{Q}\mathbf{x} + \mathbf{c}^t\mathbf{x}$. If $S = \emptyset$, the program is called infeasible and there is no solution to the program.

We determine $\{x_{lt_j}\}$ by quadratic programming with a simple, suitable objective function ($\mathbf{Q} = \mathbf{I}$; $\mathbf{c} = \mathbf{0}$),

$$\text{Minimize } \sum_{\text{all } x} x_{lt_j}^2 \quad (8)$$

subject to the constraints described in the previous section. The objective function has equal weight on each parameter and tries to keep the parameters as small as possible without influencing their sign; there is no a priori condition such as that a specific $e_{lt_j}(r_{ij})$ should be attractive or repulsive. The size and sign of the resulting parameters are thoroughly dependent on the nonnatives generated.

Since the number of nonnatives is huge, the whole set of inequalities cannot be included all at once. After a certain number of violated inequalities of the current parameters (or solution) are found, the system is solved to obtain a new set of parameters. Only inequalities with small slack (i.e., the distance between the point of the current solution and the hyperplane of the inequality in the parameter space) are kept for the next cycle. For each cycle, we carry out several energy minimizations starting with a randomly chosen nonnative or the native until it converges; after every 200 steps, the inequality of the current conformation is added to the system if it is violated. The minimizations are done in the continuous state conformation space, or in terms of ϕ s and ψ s. This is repeated until the current solution reaches a real solution (i.e., not a single violated inequality is found). Please note that here the potential function itself is not the objective to be optimized. The method is employed to find one of the solutions, $\{x_{lt_j}\}$, in the feasible region, and the only requirement on the system is that each nonnative should have a higher energy than the native, as described in the previous section.

RESULTS

After around 3,000 minimization trials, we have successfully determined the parameters for 1ubi (Figure 3) for the threaded, randomized, and perturbed (t/r/p) nonnative conformations (Figure 4), and the

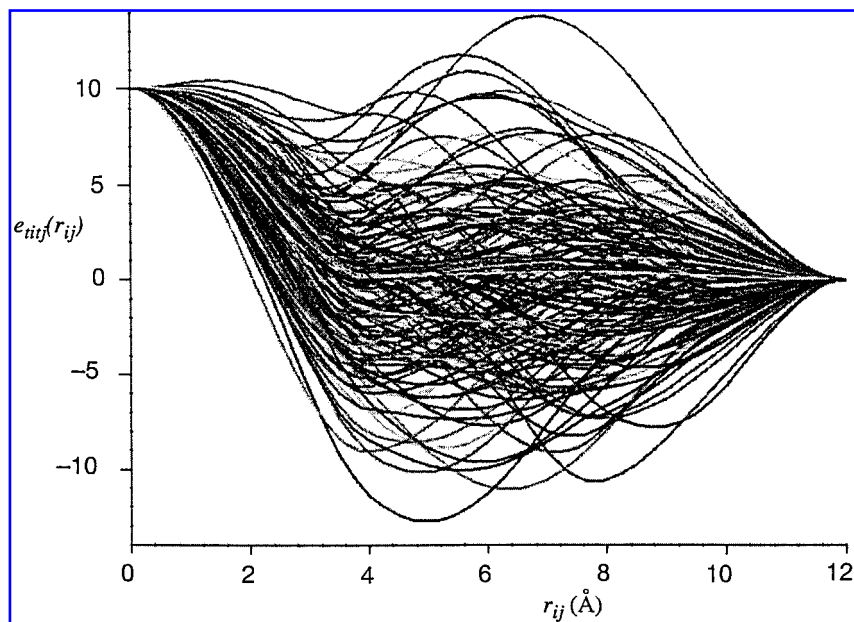


FIG. 3. The $e_{lt_j}(r_{ij})$ trained for 1ubi.

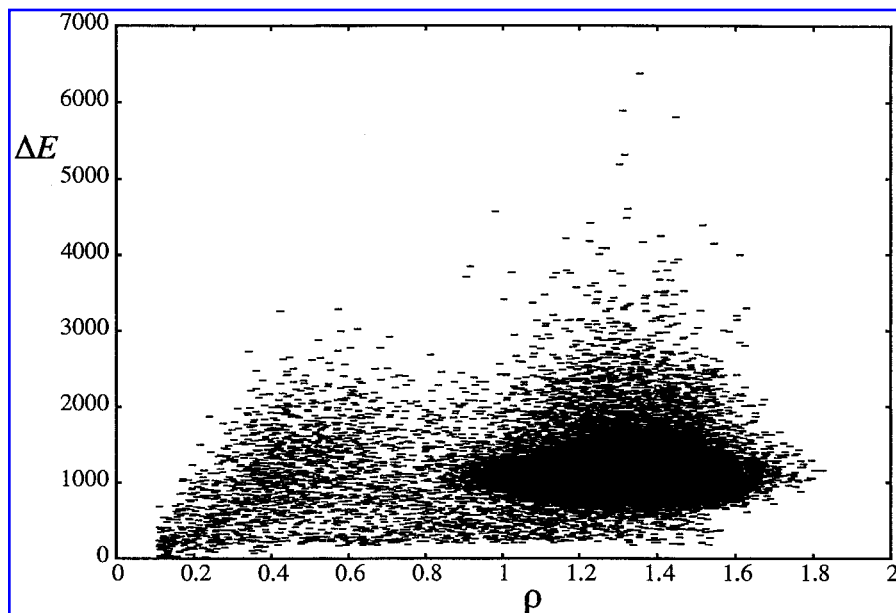


FIG. 4. The distribution of 1ubi's nonnatives ($t/r/p$) in ΔE and ρ . The average of ΔE is 1132.9 ± 344.8 , that of ρ is 1.27 ± 0.24 , the correlation between them is -0.100 .

energy-minimized ones ($\min E$, Figure 5), too. Since conformation space has many dimensions and the least upper bound on $|\partial E/\partial \phi|$ is large, it is not feasible to prove that the energy of the native is truly at the global minimum, but no inequalities of $t/r/p$ conformations are violated by the parameters obtained, and so far we have found no $\min E$ conformation of negative ΔE ($= E_{\text{nonnat}} - E_{\text{nat}}$). On the other hand, the potential trained for $t/r/p$ nonnatives only (Figures 6 and 7) does not work well. All the $\min E$ nonnatives have large negative ΔE (Figure 7). The energy minimization starting from the native converges to a conformation (or $\min E$ native) far away from the native (ρ to the native is 0.400). Apparently, by the $t/r/p$ -trained potential the native is not even at a local minimum, let alone at the global minimum.

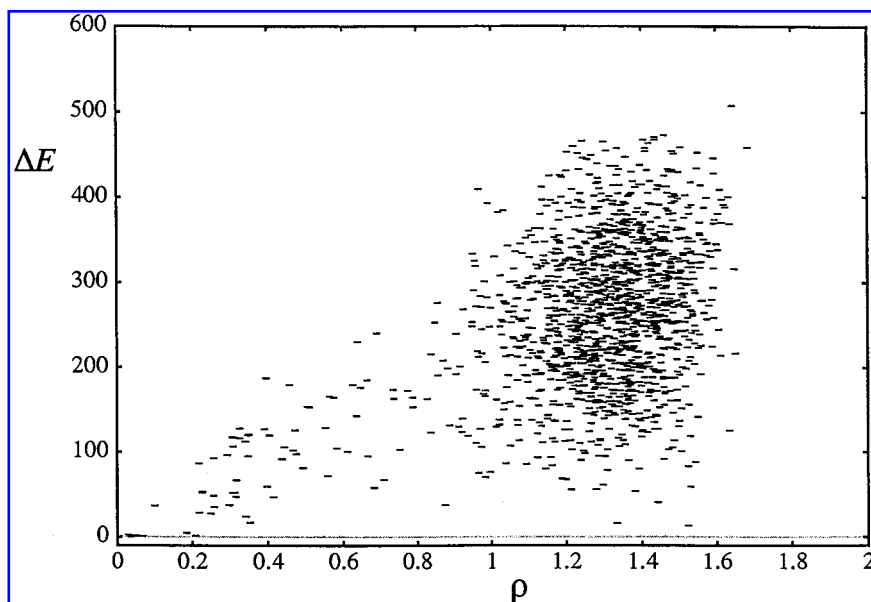


FIG. 5. The distribution of 1ubi's nonnatives ($\min E$) in ΔE and ρ . The average of ΔE is 256.5 ± 92.6 , that of ρ is 1.27 ± 0.25 , the correlation between them is 0.477 .

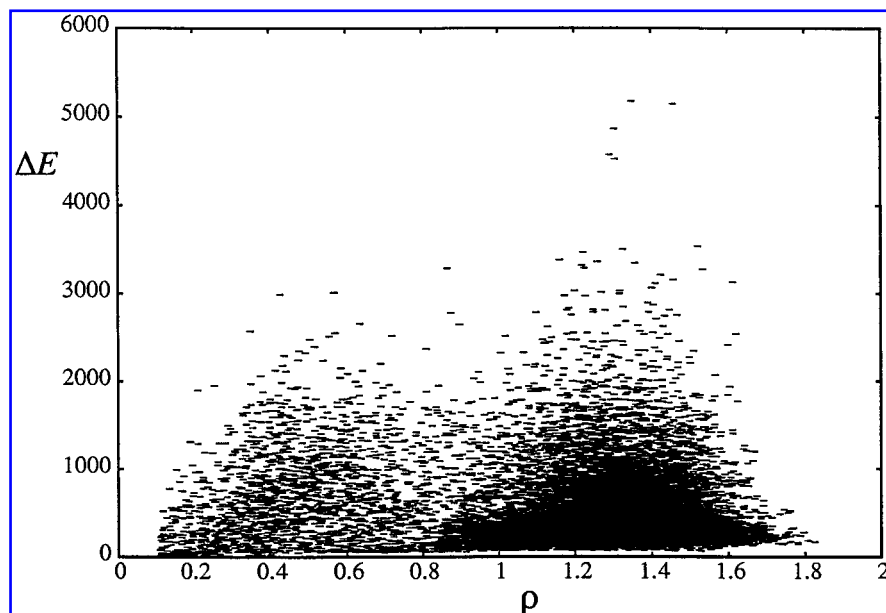


FIG. 6. The distribution of Iubi's nonnatives (*t/r/p*) in ΔE and ρ , by a potential trained for *t/r/p* nonnatives only. The average of ΔE is 379.9 ± 303.4 , that of ρ is 1.27 ± 0.24 , and the correlation between them is -0.151 .

Several minimizations were observed to converge from a fairly different conformation (ρ to the native was around 0.5) to the native. It will be interesting to compare the radius of convergence for the native and those for the nonnative minima to test the hypothesis of Shortle *et al.* (1998) on this continuous state model, namely that the native minimum is broader than nonnative minima.

The parameters obtained were applied to another protein, 1bt0, which is not listed in Table 1, having 62% sequence identity, the same types of residues, and 0.7 Å RMSD to Iubi. There are no *t/r/p* nonnatives having negative ΔE (Figure 8). Some *minE* nonnatives have negative but small ΔE (Figure 9). The rest have large negative ΔE , yet all of them have higher energies than the *minE* native, which is so similar

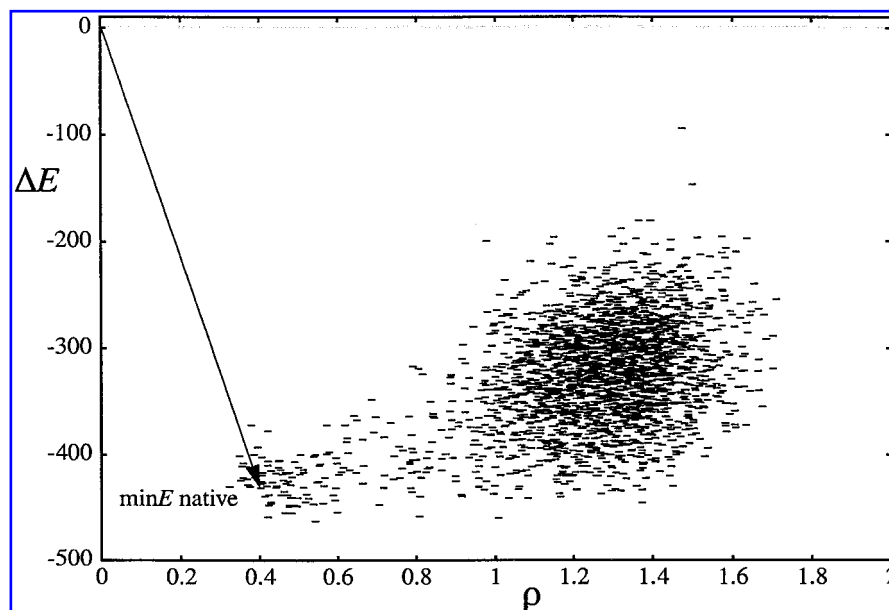


FIG. 7. The distribution of Iubi's nonnatives (*minE*) in ΔE and ρ , by a potential trained for *t/r/p* nonnatives only. The average of ΔE is -339.7 ± 57.0 , that of ρ is 1.24 ± 0.22 , and the correlation between them is -0.424 .

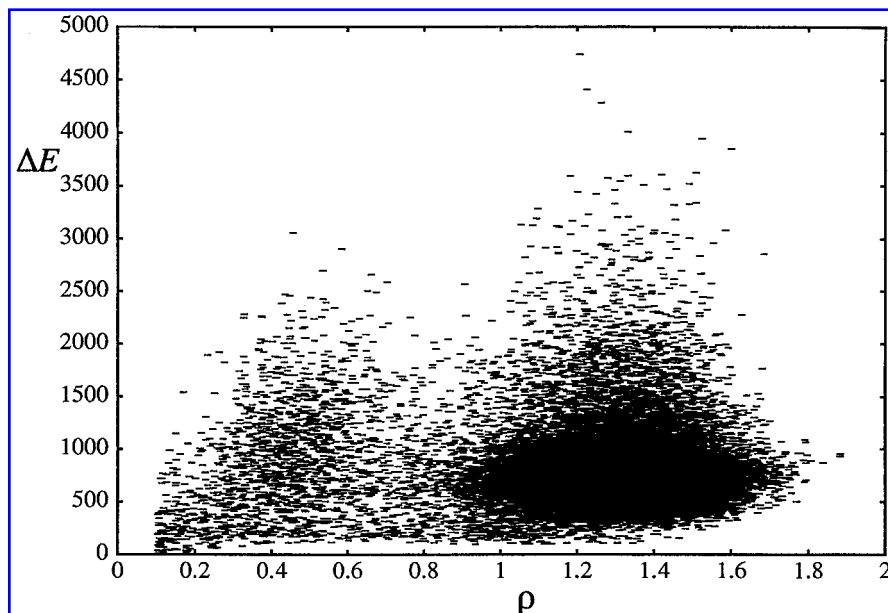


FIG. 8. The distribution of 1bt0's nonnatives ($t/r/p$) in ΔE and ρ . The average of ΔE is 767.6 ± 320.3 , that of ρ is 1.27 ± 0.24 , and the correlation between them is -0.013 .

to the native ($\rho = 0.129$) as to be regarded as one of the natives (although the ρ slightly violates the set tolerance for being one of the natives, 0.10; see *Constraints* section). So the native 1bt0 is not exactly at, but is close to, a local minimum. Besides, this local minimum seems to be the global minimum. Thus, the parameters adjusted for 1ubi are also good for another similar protein, 1bt0.

There is a possibility, however, that the potentials obtained are overfitted to 1ubi and that the potential does not show general properties of protein structures anymore. To check for overfitting, we carried out a simple test. The potential was applied to fifteen residue, right- and left-handed helices (Table 2). The standard geometry conformations are generated using dihedral angles of $(-64, -40)$ for the right-handed,

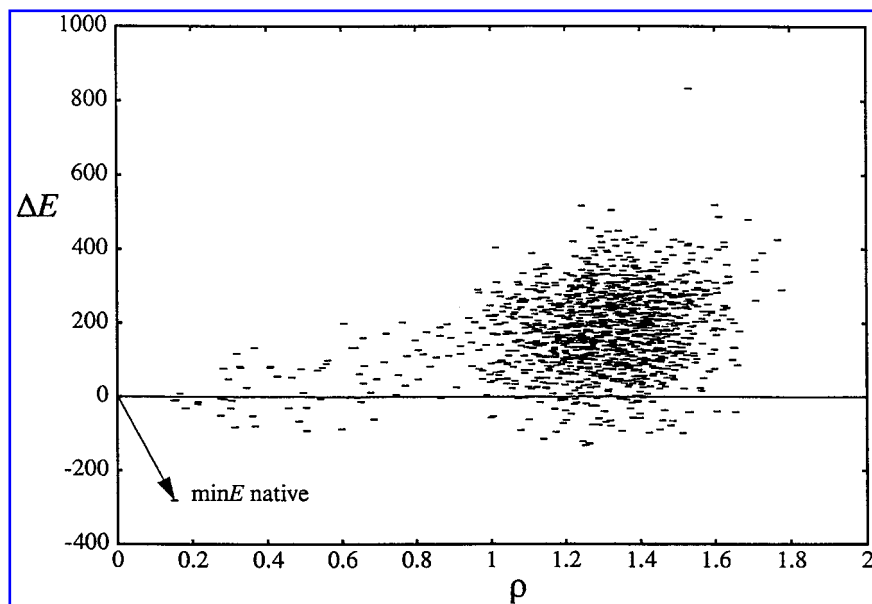


FIG. 9. The distribution of 1bt0's nonnatives ($\min E$) in ΔE and ρ . The average of ΔE is 177.1 ± 110.6 , that of ρ is 1.26 ± 0.23 , the correlation between them is 0.319 .

TABLE 2. THE ENERGIES OF 15 RESIDUE, RIGHT- AND LEFT-HANDED HELICES. E AND $\min E$ ARE THE ENERGY BEFORE AND AFTER MINIMIZATION, RESPECTIVELY THE SEQUENCE OF POLYLS IS LLSSLLSLLSSLLSL

		E	$\min E$	ρ
polyA	r-handed	-110.0	-118.3	0.110
	l-h	-56.0	-106.3	0.570
polyS	r-h	-314.5	-323.9	0.019
	l-h	-252.1	-277.2	0.131
polyLS	r-h	-212.6	-232.1	0.041
	l-h	-151.4	-178.2	0.239

and (64, 40) for the left-handed. The energies are calculated before and after the energy minimization of each helix. In all cases, the right-handed have a lower energy both before and after minimization and smaller conformational change than the left-handed. Thus, the potential obtained gives greater stability to the right-handed helices than to the left-handed, as it should. While this result is certainly not revolutionary, it is actually not a trivial feature of potential functions. The balance between the two helices is actually a fine difference in sidechain-sidechain and sidechain-backbone distances between residues that are represented in enough detail to have the correct chirality. This has been learned over the course of numerous conformational comparisons which were never specifically aimed at this distinction. If the training is carried out with only threaded nonnatives, all training helices are right-handed and the resulting potential does not clearly favor either form. In fact, neither the right- nor the left-handed helix is near a local minimum of that potential (data not shown).

Even though the potential has been trained with unprecedented rigor over a great range of nonnative conformations, only a single native sequence was used. Naturally, we were curious to see whether it would behave at all reasonably for unrelated sequences. We selected eight proteins without chain breaks, disulfide bonds, or ligands from Table 1 and checked their energies and ranks among $t/r/p$ conformations (Table 3). None is ranked first, and while all at least have (favorable) negative Z-scores, some are not very impressive. One problem is that ubiquitin has no Cys or Trp residues (nor does 1bt0), so all such interactions in the potential have only a repulsive core at short distances (b_0) and are otherwise zero at longer distances (see Discussion section). Out of this test set, only 1div and 1bv1 also lack Cys and Trp; all those having native Z-scores greater than -0.5 contain such residues. It is obvious that the potential should be trained for a set of proteins to be more generic.

TABLE 3. THE ENERGIES AND Z-SCORES OF THE NATIVES AND CONFORMATIONS RANKED FIRST BY THE POTENTIAL FUNCTION TRAINED FOR IUBI

PDB code	Residues	Class	Rank-1 conformation		Native	
			ΔE^a	Z-score	Z-score	Rank ^b
1lfc	131	β	-191.56	-1.64	-1.50	129
2end	137	α	-265.08	-1.70	-1.29	255
1div	149	$\alpha + \beta$	-366.47	-1.55	-1.05	1,439
2i1b	153	β	-963.01	-1.68	-0.53	14,597
1bv1	159	$\alpha + \beta$	-447.83	-1.70	-1.19	517
119l	162	$\alpha + \beta$	-473.81	-1.52	-1.01	1,426
1amm	174	β	-875.50	-1.57	-0.74	7,520
1akz	223	α/β	-974.80	-1.49	-0.82	6,361

^a $\Delta E = E - E_{nat}$.

^bThere are approximately 43,000 total $t/r/p$ nonnative conformations in each case.

DISCUSSION

Model

We fitted PDB structures to continuous state models. The fitting was needed because the rebuilt conformation would be fairly, or sometimes very, different from the PDB structure if we simply used the values of ϕ s and ψ s in the PDB structure. Moderate deviation of ω s from 180° and/or *cis*-Pro in the middle of a chain mainly contribute to the difference. The deviations in other bond angles and lengths do to a certain degree, too. Minimizing a penalty function P is a simple and complete way to solve this problem.

It should be noted that there are numerous continuous state models (or sets of dihedral angles) whose ρ to the original PDB structure is small. The fitting method will find just one of them that may not be of the smallest possible ρ to a given PDB structure, just as the discrete state method of Park and Levitt (1995) does. Different models of small ρ can be obtained by, for instance, assigning different weights to terms in the penalty function P . However, those models are all very close to each other and to the PDB structure. The model used here, therefore, can be safely used as the representative of the original structure in the continuous space. The difference in ρ , and hence RMSD, between the native model and the PDB structure is small compared to the resolution of the PDB, which is ordinarily larger than 1 Å.

For the same reason, it is allowable to switch the native model during the training to an energy-minimized one, if it is also one of the “natives.” It is often the case that energy minimization starting from the native converges quickly to a slightly different conformation. In this case, we replace the initial native by the energy-minimized, so as to complete the training quickly.

As the native, Iubi was chosen in this study. Among 313 proteins used, this protein is the smallest compact one without chain breaks, disulfide bonds, or ligands. This condition is necessary for the native, because our simplified model cannot describe the existence of these, and our simple form of potential does not have the term of the energies for disulfide bridges or the interactions between proteins and ligands; for nonnatives, it matters but not so much as for the native. We selected the smallest for the native simply to save computation time. We used the same standard (except regarding the size) to choose the test proteins for Table 3.

Potential function

We have shown that a pairwise-type potential function can identify the native conformation as the (near-)global minimum compared to other nonnatives, including energy-minimized ones. Although our potential is not optimized to maximize the energy gaps between the native and others (e.g., Mirny and Shakhnovich, 1996; Klimov and Thirumalai, 1998), and therefore the Z-scores are not large (-3.29 in Figures 4 and 5 combined; -2.40 in Figures 8 and 9 combined), the native has a lower energy than other local minima.

We used gaplessly threaded, randomized, and perturbed (or t/r/p) conformations as starting points of energy minimization in order to obtain low-energy nonnatives. We did not use gapped threading (e.g., Bryant and Lawrence, 1993; Lathrop and Smith, 1996; Akutsu and Tashimo, 1998; Crawford, 1999). Conformations by gapped threading are expected to have a lower energy than ones by gapless threading, but still there is no guarantee that they are at local minima. While (gapless-)t/r/p/ nonnatives distribute widely along ΔE (on the order of thousands) and have small correlation between ΔE and ρ (Figures 4 and 8), $\min E$ nonnatives starting from (gapless-)t/r/p distribute along ΔE on the order of hundreds and have high correlation between ΔE and ρ , covering a wide range of conformation space (Figures 5 and 9). This suggests that, although the energy surface is rugged, its local minima are somewhat bounded. We would be surprised, therefore, if $\min E$ nonnatives from gapped-threaded conformations have much lower ΔE than other $\min E$ nonnatives.

Vendruscolo and Domany (1998) showed that pairwise contact potentials for contact maps were unable to assign to all the nonnatives a higher energy than that of the native. Domany and coworkers supported that even in a case where a contact energy is a good approximation of the real one, by using a Lennard-Jones “true” potential and its approximation (Clementi *et al.*, 1999). The conformational model and potentials used by them are all-or-none or not fully distance dependent (only the amplitudes of preset distance-dependent energy terms can be changed), while those used here are distance dependent (each $e_{i_l i_j}(r_{ij})$ can be in various forms, attractive or repulsive). Although differences in their models and ours preclude a direct comparison, our results are opposite to theirs, i.e., a trained potential function can assign the native to the lowest energy.

The potential function also does well for another protein, 1bt0; the *minE* native is close to the native and has a lower energy than *minE* nonnatives. Moving further away from the training sequence, 1ubi produces much more modest results, as shown in Table 3. This is a common problem with empirical potentials, whether they are trained over a narrow range of sequences and a broad field of structures (Figure 1, bar (c)), or over a broad selection of sequences and a restricted collection of structures (Figure 1, bar (a)). For example, the simple hydrophobic potential of Huang *et al.* (1996) works well for many sequences challenged by threaded nonnative structures, but it fails specifically on ubiquitin (1ubq, which is extremely similar to 1ubi) when tested on perturbed structures generated by molecular dynamics.

Training method and energy surface

As described in the previous section, the energy surface by the potential trained by our method is more or less bottom-bounded (Figure 10a). On the other hand, the energy surface by a potential trained only by threading may be rather top-bounded (Figure 10b) or neither (Figure 10c), as, once again, threaded conformations tend to be away from local minima; the trained potential does not satisfy the essential requirement that the native should have a lower energy than nonnatives. In our settings, the *t/r/p*-trained potential seems bottom bounded (Figures 6 and 7), but we suspect this tendency is smaller when the *t*-trained potential (not trained for the randomized or perturbed conformations) is used (data not shown).

The bottom-bounded tendency along with high correlation between ΔE and ρ (or any conformational similarity metric) is a desirable feature for conformational search. A search strategy focusing on the local minima (e.g., the CGU method by Dill and coworkers [1997]) should work well on an energy surface with

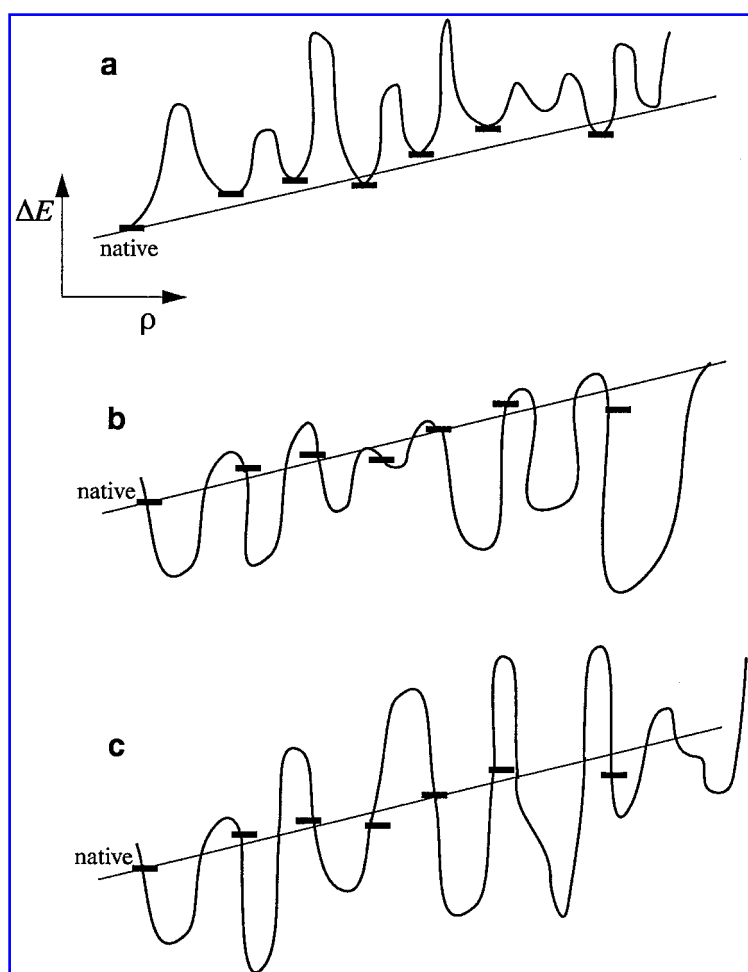


FIG. 10. The schematic diagrams of the energy surface. The bars signify the native or nonnatives used for potential training. The curves represent the energy surface: (a) a bottom-bounded energy surface, (b) a top-bounded energy surface, (c) a neither top- nor bottom-bounded, just rugged energy surface.

such a feature. It is clear from Figures 4 and 5 that the energy surface by our trained potential is an instance of such an energy surface, and that the one obtained from threading is not. Monte Carlo conformation search is currently under way where each step is followed by energy minimization in order to hop among local minima.

Drawbacks and conclusion

There are some limitations or inconveniences to this approach. Some potential energies in Figure 3, especially repulsive ones, have peculiar shapes. This is a result of the choice of basis functions. It is difficult to find a small number of basis functions which are orthogonal in the interaction range, can span a wide range of r_{ij} , and yet in any linear combination have positive values at short distances and gradually reach zero at long distances. This condition would be fairly easily realized if a large number of basis functions were employed, but with an even larger number of adjustable parameters, $\{x_{llt_j}\}$, and increased risk of overfitting. We have tried several basis function sets, and the one in this paper, although not mutually orthogonal, works best so far.

As described in the Methods Section, quadratic programming finds the unique point in the feasible region. There is no guarantee, however, that the solution obtained is the best (or a good enough) one unless the size of the feasible region is well restricted and small enough so that any point in the feasible region can be a good solution. The objective function $\sum_{\text{all } x} x_{llt_j}^2$ itself does not have any physical meaning and is not a specific constraint on any particular atom pair's interaction. We need enough constraints (or inequalities) for each atom pair to avoid loose conditioning and then we need a particular solution.

Some $e_{t_i t_j}(r_{ij})$ are untrained if the atom pairs of t_i and t_j do not exist in the native. Since those $e_{t_i t_j}(r_{ij})$ have no constraints, quadratic programming automatically sets their parameters $x_{1t_i t_j}$, $x_{2t_i t_j}$, and $x_{3t_i t_j}$ at 0 to minimize the objective function. So those $e_{t_i t_j}(r_{ij})$'s are determined, but not trained, and may not work well. It is natural and not surprising that the $e_{t_i t_j}(r_{ij})$'s trained for 1ubi works well for a similar protein, 1bt0, but not for larger, unrelated proteins which have numerous $t_i t_j$ pairs not present in 1ubi. The parameters, therefore, should be adjusted for a couple of proteins to make the potential function more generic. For that purpose, it seems important to find the smallest set of proteins with which any (or some specific) potential function trained can work well for the largest possible number of other proteins. It seems the current potential can be better and more easily trained (i.e., so that the $\min E$ native = the native) by adding some constraint inequalities, without changing the general property of the current potential or resulting in an infeasible system. It may be not so difficult to train for a couple of extra proteins, too. Training for a different protein, training for multiple proteins, and comparison of those potentials are currently being carried out in order to check for feasibility and overfit.

The successful determination of the parameters suggests that the continuous state model and potential functions used here can be used as a realistic yet simple model of proteins. Our settings might be a little too relaxed. For example, the functions do not have constraints among "natives," since nonnatives with ρ of less than 0.1 from the native are regarded as natives and are excluded from the inequality set. Although our minimization data show that the ΔE 's of those natives are small positive values and satisfy the condition on ΔE automatically, there still is a possibility that some natives have large, positive (or even negative) ΔE . We are testing an additional condition that ΔE 's among natives be minimized.

In summary, we have shown a new way to obtain realistic yet simple protein potential energy functions in a continuous conformational space, which is neither Boltzmann nor knowledge based. The potential obtained satisfies our condition; the potential so far does not ensure that many native proteins have lower energies than a relatively small number of nonnatives. Instead, the potential guarantees that the native conformation for one sequence has a lower energy than (almost) all the nonnatives, including low-energy nonnatives all over the conformation space, and seems to work in the same way for another protein, too. Besides, there seems room to train the potential for multiple proteins. Training for multiple proteins is currently underway.

ACKNOWLEDGMENTS

This work is supported by NSF (DBI-9614074) and NIH (GM59097-01) and is in debt to all the PDB contributors and administrators. The support by RECOMB 2000 is also highly appreciated. The authors are grateful to anonymous referees for their critical comments and to Dr. V. N. Maiorov, as some programs used here are the revised versions of his originals.

REFERENCES

- Akutsu, T., and Tashimo, H. 1998. Linear programming based approach to the derivation of a contact potential for protein threading. *Proc. Pac. Symp. Biocomp.* 413–424.
- Bazaraa, M.S., Sherali, H.D., and Shetty, C. M. 1993. *Nonlinear Programming: Theory and Algorithms*, 2nd ed., John Wiley and Sons, Inc., New York.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The protein data bank. *Nucleic Acids Res.*, 28, 235–242.
- Bernstein, F.C., Koetzle, T.F., Williams, G.G.B., Meyer, F.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.
- Bowie, J.U., Lüthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170.
- Bryant, S.H., and Lawrence, C.E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16, 92–112.
- Clementi, C., Vendruscolo, M., Maritan, A., and Domany, E. 1999. Folding Lennard-Jones proteins by a contact potential. *Proteins* 37, 544–553.
- Crawford, O.H. 1999. A fast, stochastic threading algorithm for proteins. *Bioinformatics* 15, 66–71.
- Crippen, G.M. 1996. Easily searched protein folding potentials. *J. Mol. Biol.* 260, 467–475.
- Crippen, G.M., and Ohkubo, Y.Z. 1998. Statistical mechanics of protein folding by exhaustive enumeration. *Proteins* 32, 425–437.
- Deutsch, J.M., and Kurosky, T. 1996. New algorithm for protein design. *Phys. Rev. Lett.* 76, 323–326.
- Dill, K.A., Phillips, A.T., and Rosen, J.B. 1997. Protein structure and energy landscape dependence on sequence using a continuous energy function. *J. Comp. Biol.* 4, 227–239.
- Gō, N. 1983. Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.* 12, 183–210.
- Godzik, A., Kolinski, A., and Skolnick, J. 1995. Are proteins ideal mixtures of amino-acids? Analysis of energy parameter sets. *Protein Sci.* 4, 2107–2117.
- Hao, M., and Scheraga, H.A. 1996. Optimizing potential functions for protein folding. *J. Phys. Chem.* 100, 14540–14548.
- Hendlich, M., Lackner, P., Weitkus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M.J. 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* 216, 167–180.
- Hestenes, M.R. 1980. *Conjugate Direction Methods in Optimization*. Springer-Verlag, New York.
- Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci.* 3, 522–524.
- Huang, E.S., Subbiah, S., and Levitt, M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252, 709–720.
- Huang, E.S., Subbiah, S., Tsai, J., and Levitt, M. 1996. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol.* 257, 716–725.
- Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.* A34, 827–828.
- Klimov, D.K., and Thirumalai, D. 1998. Cooperativity in protein folding: From lattice models with sidechains to real proteins. *Fold. Design* 3, 127–139.
- Lathrop, R.H., and Smith, T.F. 1996. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* 255, 641–665.
- Maiorov, V.N., and Crippen, G.M. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227, 876–888.
- Maiorov, V.N., and Crippen, G.M. 1995. Size independent comparison of protein three-dimensional structures. *Proteins* 22, 273–283.
- Mirny, L.A., and Shakhnovich, E.I. 1996. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* 264, 1164–1179.
- Miyazawa, S., and Jernigan, R.L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18, 534–552.
- Miyazawa, S., and Jernigan, R.L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256, 623–644.
- Momany, F.A., McGuire, R.F., Burgess, A.W., and Scheraga, H.A. 1975. The mean geometry of the peptide unit from crystal structure data. *Biochim. Biophys. Acta* 359, 298–302.
- Nishikawa, K., and Matsuo, Y. 1993. Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* 6, 811–820.
- Park, B.H., and Levitt, M. 1995. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 249, 493–507.
- Park, B., and Levitt, M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258, 367–392.

- Park, B., Huang, E.S., and Levitt, M. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266, 831–846.
- Schulz, G.E., and Schirmer, R.H. 1979. *Principles of Protein Structure*. Springer-Verlag, New York.
- Shortle, D., Simons, K.T., and Baker, D. 1998. Clustering of low-energy conformations near the native structures of small protein. *Proc. Natl. Acad. Sci. USA* 95, 1158–1162.
- Thomas, P.D., and Dill, K.A. 1996a. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* 257, 457–469.
- Thomas, P.D., and Dill, K.A. 1996b. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA* 93, 11628–11633.
- Torda, A.E. 1997. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* 7, 200–205.
- Vajda, S., Sippl, M., and Novotny, J. 1997. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* 7, 222–228.
- Vendruscolo, M., and Domany, E. 1998. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* 109, 11101–11108.
- Wang, Y., Zhang, H., Li, W., and Scott, R.A. 1995. Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl. Acad. Sci. USA* 92, 709–713.
- Yue, K., and Dill, K.A. 1995. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA* 92, 146–150.

Address correspondence to:

Gordon M. Crippen
College of Pharmacy
University of Michigan
Ann Arbor, MI 48109-1065

E-mail: gcrippen@umich.edu

This article has been cited by:

1. R. Rajgaria, S. R. McAllister, C. A. Floudas. 2008. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins: Structure, Function, and Bioinformatics* **70**:3, 950-970. [[CrossRef](#)]
2. Gordon M. Crippen. 2005. Recognizing protein folds by cluster distance geometry. *Proteins: Structure, Function, and Bioinformatics* **60**:1, 82-89. [[CrossRef](#)]