

# Determining Contact Energy Function for Continuous State Models of Globular Protein Conformations

Y. Zenmei Ohkubo  
College of Pharmacy  
University of Michigan, Ann Arbor  
MI 48109-1065, USA  
zenmei@umich.edu

Gordon M. Crippen  
College of Pharmacy  
University of Michigan, Ann Arbor  
MI 48109-1065, USA  
gcrippen@umich.edu

## 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 results show that the potential can be trained to discriminate between the native conformation of one protein as the (near-)global minimum, and other nonnatives, including energy-minimized ones (or local minima). This potential function is able to identify the native conformation of another protein, too.

## 1. Introduction

To obtain a model of foldable proteins, in order to study the nature of protein folding, there are two approaches. The one approach is, given "true" potentials (which are supposed to be true), select amino acid sequences which have a unique conformation (or structural model) of the lowest energy by the given potentials (e.g., [36, 8]). The good thing about this approach is that there is no worry about obtaining the potential of protein folding, because it is known *a priori*. The sequence search is not time-consuming if the conformational space has a tractable size. The foldable sequences and conformations obtained as the native state of the sequences, however, may not be protein-like.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB 2000 Tokyo Japan USA  
Copyright ACM 2000 1-58113-186-0/00/04 \$5 00

The other approach is, given "true" protein conformations (which are known native structures of proteins), train adjustable potentials for the given conformations so that the native conformation for a sequence is the most stable, or has a lower energy than nonnative conformations (which, for instance, may include native conformations for different sequences). Protein sequences and conformations used are real, existing ones, or at least models adequately close to real ones. The main concern in this approach, therefore, is how valid the trained potentials are.

The common way to obtain potentials and check the validity is called the "recognition problem". This problem, to choose the native fold of a given protein sequence from a given zoo of conformations under the guarantee that the native fold is in the zoo, is a restricted, conditional version of the long-standing protein folding problem. Numerous approaches to the recognition problem have been carried out (e.g., [23, 12, 20, 5, 15, 31, 24]). There are comparisons of those methods (e.g., [10, 35, 27, 28]), and reviews (e.g., [32, 33] and references therein).

Although these methods vary in their assumptions, functional forms and parameters, or conformational space, one of the common ways to train (or check) the potentials is to utilize native conformations of different sequences as nonnative states of a given sequence ("threading" method, [12]). The potentials obtained should give a lower energy to the native conformation than any other nonnatives. The better potentials are expected to satisfy this condition for a larger number of sequences, because the real, unknown free energy governs every protein folding (except for kinetically determined folds).

Threading alone, however, may be too loose a condition for protein folding. 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 co-workers pointed out these shortcomings and took a challenging approach; they generated nonnatives around the native by molecular dynamics, and checked whether the native conformation had a lower energy than the low-energy nonnatives by their potential [16]. This condition is much more stringent than threading training, as nonnatives by threading are at relatively high ener-

1191	1531	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	1bel	1bea	1bei	1beo	1bf8	1bfg	1bgf
1bkf	1ble	1bol	1bor	1br0	1brf	1bsn	1btn	1buz	1bvi	1bvh	1bw3	1bxa	1bym	1c25	1c52
1c5a	1cby	1cdb	1cdi	1cex	1cfb	1cfe	1cfh	1chd	1chl	1cid	1ctj	1cto	1cur	1cyo	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	1ifc	1ife
1irl	1jer	1jli	1jpc	1juk	1jvr	1kbs	1kid	1knb	1kpf	1krt	1ksr	1kte	1kuh	1lba	1lbu
1lcl	1leb	1lit	1lki	1lou	1lrv	1mai	1mak	1mb1	1mbh	1mbj	1mrj	1msc	1msi	1mup	1mut
1mzm	1ngr	1nkl	1nkr	1nls	1noe	1nox	1npg	1nxb	1ocp	1ois	1opd	1opr	1orc	1pce	1pdo
1pex	1pft	1pih	1pkp	1plc	1pne	1poa	1poc	1pou	1ppn	1ppt	1put	1qyp	1ra9	1rcf	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	1wiu	1wkt	1xnb	1ycc	1yua	1yub
1zaq	1zin	1zug	1zwa	1zxx	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	2ilb	2igd	2ilk	2lbd	2lfb	2mcm	2nef	2new	2pac	2phy	2pii	2pth	2pt1
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							

**TABLE 1:** 313 PDB entries chosen from the 25% list of the December 1998 release of PDB\_Select [14].

gies (see Fig. 1 in [34]). Levitt and co-workers found their potentials worked well, 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 it 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.

This procedure is similar to that of Crippen [7]. 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 contact energy function which depends on contact distance and contact atom types in a continuous internal coordinate space under the conditions of fixed bond lengths and angles. The potentials are not based on Miyazawa & Jernigan's quasi-chemical approximation [23, 24], empirical knowledge, or any specific assumption, such as setting nonnative interactions at neutral in the  $\bar{G}_0$  model [9]. 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, contin-

uous and additive functions, and then adjust the parameters by the procedure outlined above.

Iterative optimization methods have been used by Hao and Scheraga [11] to maximize the Boltzmann probability of the native conformations. Mirny and Shakhnovich [22] maximized the negative of the harmonic mean of Z scores [3], which express how far away the energy of natives are from average energy of the nonnatives in terms of standard deviation units. The major difference between these approaches and ours is that the former tried to maximize the energy difference between the native and a relatively limited set of nonnatives, while our objective is to put the native at an apparent global minimum.

## 2. Methods

### 2.1. Continuous 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 (Dec. 1998 release, [14]). To reduce the size of the conformation space, we fitted each of the PDB structures [2] to a standard geometry polypeptide model having all *trans* peptide bonds. The fitted model consists of main chain heavy atoms and  $C_\beta$ s (pseudo- $C_\beta$  for Gly) with the standard values of bond lengths and angles [29]; the  $(\phi, \psi)$  values are any real numbers ranging from  $-180.0^\circ$  to  $180.0^\circ$ . This model is a simplified version of that for ECEPP [25], or a continuous version of the model of Park & Levitt [26]. The average RMSD [17] between a PDB structure and the fitted model is less than 0.5 Å.

## 2.2. Fitting a PDB structure to a continuous model

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

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

where  $i, j, k$  signify any main chain heavy atoms or  $C_{\beta}$  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 [29] of the  $i$  and  $j$  pair;  $\theta_{ijk}$  is the bond angle of any bonded atom triple  $i$ - $j$ - $k$ ;  $\theta_0$  is the standard bond angle of the corresponding atoms;  $\omega_i$  is the  $i$ -th peptide bond dihedral angle of the conformation;  $\pi$  is set at the standard dihedral angle of  $\omega$  ( $trans=180^\circ$ );  $c_i$  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  $\omega$  dihedral angle:

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

where  $x_{C_i^\alpha}$ , for instance, signifies the  $x$  coordinate of the  $i$ -th  $C_{\alpha}$ , 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  $\omega$  dihedral angle at exactly  $180^\circ$ . Any *cis*-Pro, therefore, is converted to *trans*-Pro.

The conjugate gradient method ([13] 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  $(\phi, \psi)$  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.

## 2.3. Native and nonnative models

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

- 36,646 threaded conformations from 313 fitted models;
- 1,000 conformations of the fitted lubi whose  $(\phi, \psi)$  pair at one randomly chosen residue is changed to an existing pair randomly chosen among 313 fitted models;
- same as in b), except two  $(\phi, \psi)$  pairs are randomized;
- same as in b), except five pairs are randomized;
- same as in b), except ten pairs are randomized;
- same as in b), except twenty pairs are randomized;
- same as in b), except all pairs are randomized;
- 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. 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-native. 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 *Constraints* section).

## 2.4. Contact energy potential

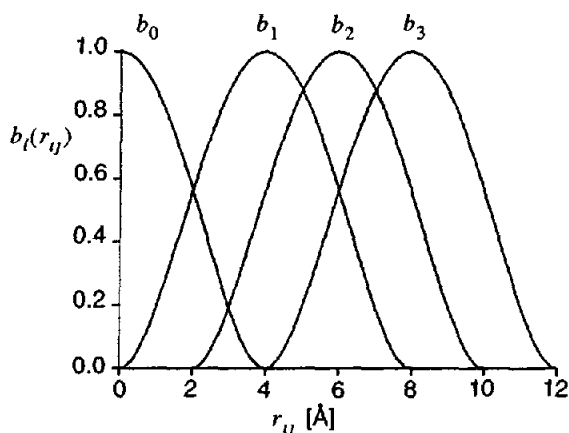
We employ a pairwise type contact energy as an energy function which can select the native conformation of a given sequence out of an assortment of conformations. The functions are atom type- and distance-dependent, and each is expressed as a linear combination of chosen basis functions. The total energy 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:

$$E = \sum_{\geq 1-4} \sum_{(i,j)} e_{t_i t_j}(r_{ij}) = \sum_{\geq 1-4} \sum_{(i,j)} \sum_{l=0}^3 x_{l,t_i} b_l(r_{ij})$$

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,

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

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  $r_{ij} = a \pm b$  (Fig. 1).  $\{x_{0,t_i t_j}\}$ , the coefficients for  $b_0$  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



**Figure 1:** 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.

-10.0 and 10.0, in order to obtain the native-discriminative energy function. That restricts each  $e_{i,t_j}(r_{ij})$  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:

$$-21.25 \leq e_{i,t_j}(r_{ij}) \leq 21.25$$

$$e_{i,t_j}(0.0) = 10.0$$

$$\left. \frac{\partial}{\partial r_{ij}} e_{i,t_j}(r_{ij}) \right|_{r_{ij}=12.0} = e_{i,t_j}(12.0) = 0.0$$

## 2.5. Number of parameters

As described above, there are three adjustable parameters for each atom type pair:  $x_{1t,t_j}$ ,  $x_{2t,t_j}$ , and  $x_{3t,t_j}$ . Since the number of atom types is nineteen (main chain heavy atoms and  $C_\beta$ 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  $3 \times 190 = 570$ .

## 2.6. Constraints

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

$$\Delta E = E_{\text{nonnat}} - E_{\text{nat}} > g,$$

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

where  $\rho$  is a size-independent metric of conformational similarity [21] between the native and nonnative;  $\rho$  equals 0 if two conforma-

tions are identical, and  $\rho$  reaches 2 when two conformations are totally dissimilar ( $\sim 1.8$  in the case of a chain of equally spaced particles, such as these protein models).  $\rho$  is about one tenth of RMSD in case of 1ub1. The proportionality between  $g$  and  $\rho$  is set for smaller  $\rho$  so that the requirement is not too strict for "neighbor" nonnatives. Ordinarily, if  $\rho$  is less than 0.3, the nonnative keeps the native's topology. As Vendruscolo & Domany [34] 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  $\rho$  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".

## 2.7. Quadratic programming for parameter adjustment

The adjustable parameters,  $\{x_{i,t_j}\}$ , are determined by quadratic programming (for reference, see [1]). 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:

$$\text{Minimize } \frac{1}{2} \mathbf{x}' \mathbf{Q} \mathbf{x} + \mathbf{c}' \mathbf{x}$$

$$\text{subject to } \mathbf{A} \mathbf{x} \geq \mathbf{b}$$

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  $S = \{\mathbf{x} \in R^n | \mathbf{A} \mathbf{x} \geq \mathbf{b}\}$  is called the feasible region of this quadratic program. Quadratic programming finds the optimal solution, the  $\mathbf{x} \in S$  which gives the minimum value of  $\frac{1}{2} \mathbf{x}' \mathbf{Q} \mathbf{x} + \mathbf{c}' \mathbf{x}$ . If  $S = \emptyset$ , the program is called infeasible and there is no solution to the program.

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

$$\text{Min } \sum_{\text{all } x} x_{i,t_j}^2$$

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 like a specific  $e_{i,t_j}(r_{ij})$  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. This is repeated until the current solution reaches a real

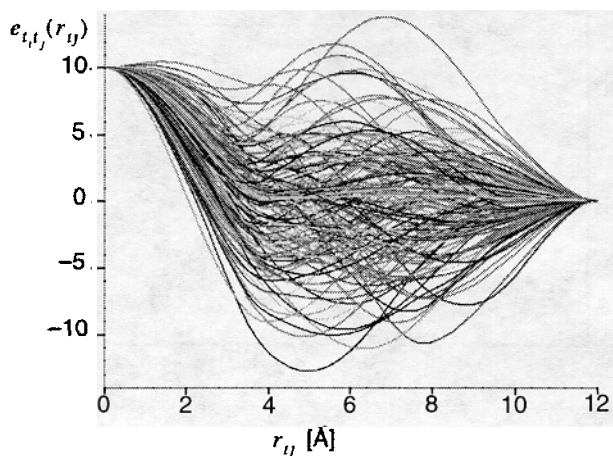


Figure 2: The  $e_{t,t_j}(r_{ij})$  trained for 1ubi.

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,  $\{x_{t,t_j}\}$ , in the feasible region and the only requirement on the system is that the nonnative should have a higher energy than the native, as described in the previous section.

### 3. Results

After around 3,000 minimization trials, we have successfully determined the parameters for 1ubi (Figs. 2) for the threaded, randomized, and perturbed ( $t/r/p$ ) nonnative conformations, and the energy-minimized ones ( $\min E$ ), too (Figs. 3a & b). 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  $\Delta E$  ( $= E_{\text{nonnat}} - E_{\text{nat}}$ ). On the other hand, the potential trained for  $t/r/p$  nonnatives only (Figs. 4a & b) does not work well. All the  $\min E$  nonnatives have large negative  $\Delta E$  (Fig. 4b). The energy minimization starting from the native converges to a conformation (or  $\min E$  native) far away from the native ( $\rho$  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.

Several minimizations were observed to converge from a fairly different conformation ( $\rho$  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 verify the hypothesis of Baker and co-workers [30] on this continuous 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 and 0.7 Å RMSD to 1ubi (Fig. 5). There are no  $t/r/p$  nonnatives having negative  $\Delta E$  (Fig. 5a). Some  $\min E$  nonnatives have negative but small  $\Delta E$  (Fig. 5b). The rest have large negative  $\Delta E$ , yet all of them have higher energies than the  $\min E$  conformation, which is

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

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 LLSLLSLLSS-LLSL.

so similar to the native ( $\rho = 0.129$ ) as to be regarded as one of the natives. So the native 1bt0 is not exactly at, but 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 protein, 1bt0

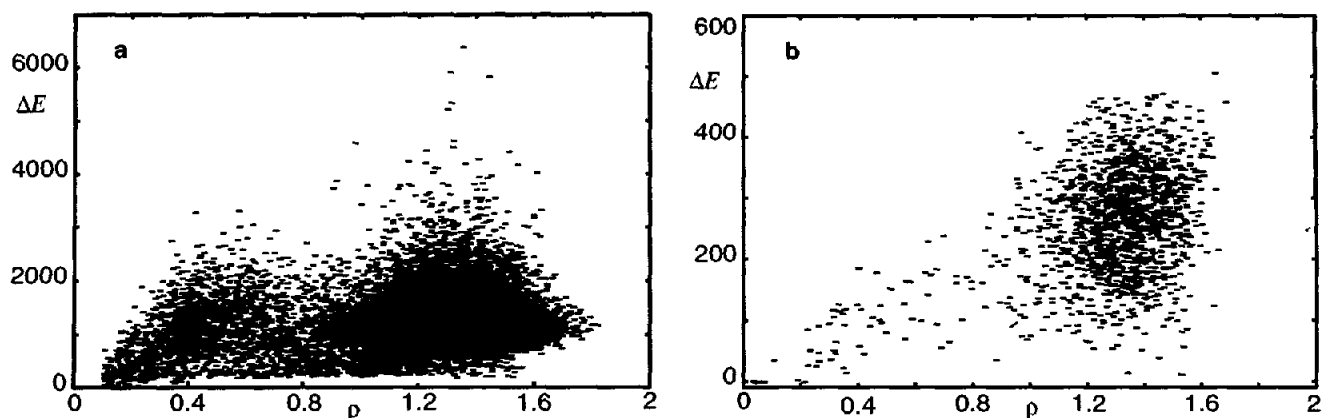
There is a possibility, however, that the potentials obtained are overfitted to 1ubi, and that the potential does not show general properties of protein structures anymore. To check for overfitting, we carried out a simple test. The potential was applied to fifteen residue, right- and left-handed helices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

## 4. Discussion

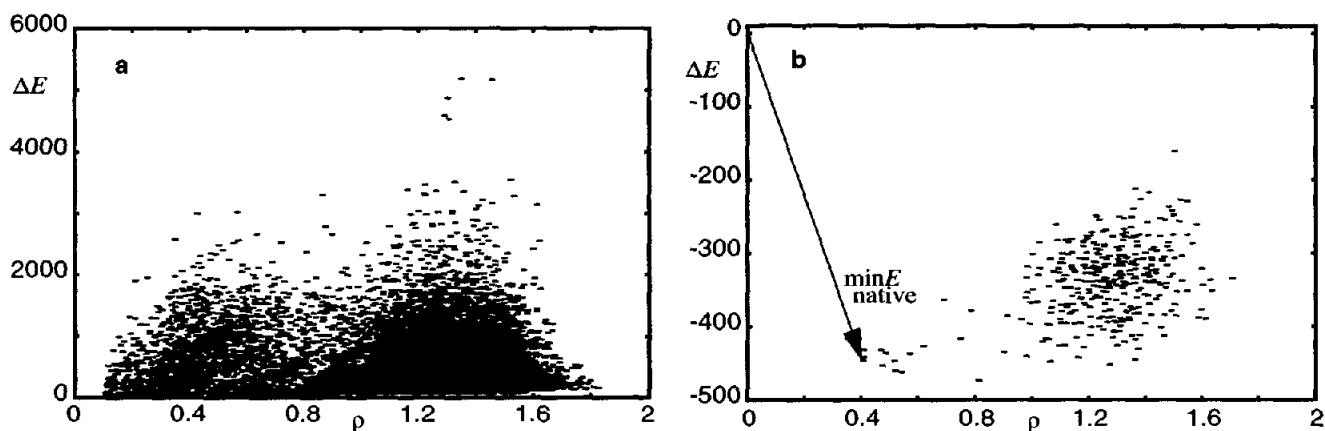
### 4.1. Model

We fitted PDB structures to continuous models. It should be noted that there are numerous continuous models (or sets of dihedral angles) whose  $\rho$  to the original PDB structure is small. The fitted model we used is just one of them. Others can be obtained by assigning different weights to terms in the penalty function  $P$ . The model used may not have the smallest  $\rho$  to the PDB structure, like the discrete model of Park & Levitt [26]. Yet it can be safely used as the representative of the original structure in the continuous space, because the models and their  $\rho$ 's are quite close to each other. Besides, the difference in  $\rho$ , and hence RMSD, between the native model and the PDB structure is small compared to the resolution of the PDB (ordinarily larger than 1Å).

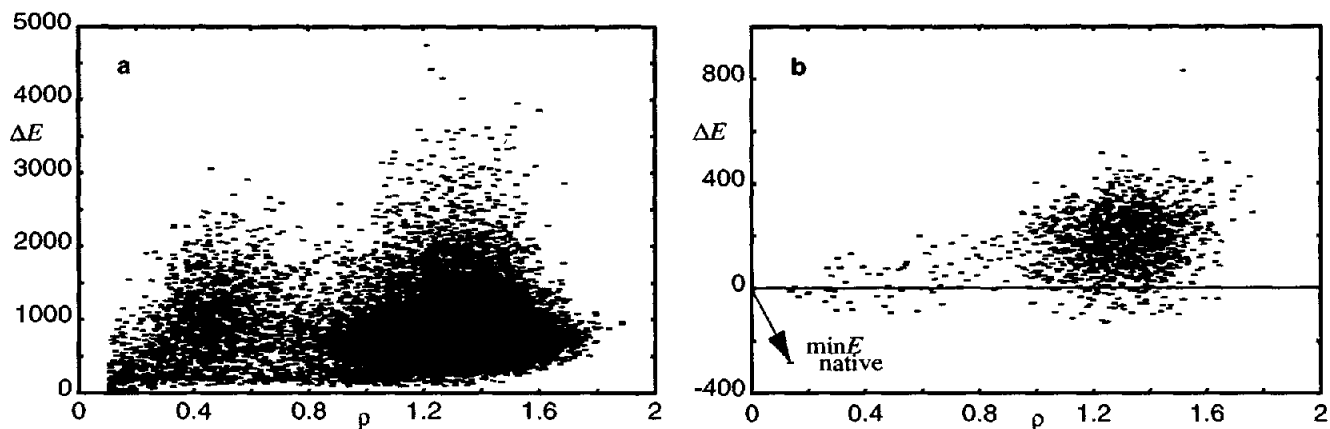
For the same reason, it is allowable to switch the native model during the training to an energy-minimized one, if it is also one of the "natives". It is often the case that energy minimization starting from the native converges quickly to a slightly different conformation. In this case, we replace the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



**Figure 3:** (a) The distribution of lubi's nonnatives ( $t/r/p$ ) in  $\Delta E$  and  $\rho$ . The average of  $\Delta E$  is  $1132.9 \pm 344.8$ , that of  $\rho$  is  $1.27 \pm 0.24$ , the correlation between them is  $-0.100$ . (b) The distribution of lubi's nonnatives ( $minE$ ) in  $\Delta E$  and  $\rho$ . The average of  $\Delta E$  is  $256.5 \pm 92.6$ , that of  $\rho$  is  $1.27 \pm 0.25$ , the correlation between them is  $0.477$ .



**Figure 4:** (a) The distribution of lubi's nonnatives ( $t/r/p$ ) in  $\Delta E$  and  $\rho$ , by potentials trained for  $t/r/p$  nonnatives only. The average of  $\Delta E$  is  $379.9 \pm 303.4$ , that of  $\rho$  is  $1.27 \pm 0.24$ , the correlation between them is  $-0.151$ . (b) The distribution of lubi's nonnatives ( $minE$ ) in  $\Delta E$  and  $\rho$ , by potentials trained for  $t/r/p$  nonnatives only. The average of  $\Delta E$  is  $-337.6 \pm 50.9$ , that of  $\rho$  is  $1.25 \pm 0.20$ , the correlation between them is  $-0.432$ .



**Figure 5:** (a) The distribution of lbt0's nonnatives ( $t/r/p$ ) in  $\Delta E$  and  $\rho$ . The average of  $\Delta E$  is  $767.6 \pm 320.3$ , that of  $\rho$  is  $1.27 \pm 0.24$ , the correlation between them is  $-0.013$ . (b) The distribution of lbt0's nonnatives ( $minE$ ) in  $\Delta E$  and  $\rho$ . The average of  $\Delta E$  is  $177.1 \pm 110.6$ , that of  $\rho$  is  $1.26 \pm 0.23$ , the correlation between them is  $0.319$ .

## 4.2. Contact potential

We have shown that a pairwise contact potential 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., [5, 22, 18]), and therefore the Z-scores are not large (-3.29 in Fig. 3; -2.40 in Fig. 5), the native has a lower energy than other local minima.

We used randomized, perturbed, and ungapped-threaded conformations as starting points of energy minimization in order to obtain low-energy nonnatives. We did not use gapped threading (e.g., [4, 19, 6]). Conformations by gapped threading are expected to have a lower energy than ones by ungapped threading, but still there is no guarantee that they are at local minima. While (ungapped-)/ $tr/p$  nonnatives distribute widely along  $\Delta E$  (the order of thousands) and have small correlation between  $\Delta E$  and  $\rho$  (Figs. 3a & 5a),  $minE$  nonnatives starting from (ungapped-)/ $tr/p$  distribute along  $\Delta E$  on the order of hundreds and have high correlation between  $\Delta E$  and  $\rho$ , covering a wide range of conformation space (Figs. 3b & 5b). This suggests that although the energy surface is rugged, its local minima are somewhat bounded. We would be surprised, therefore, if  $minE$  nonnatives from gapped-threaded conformations have much lower  $\Delta E$  than other  $minEs$ .

Vendruscolo & Domany [34] showed that pairwise contact potentials for contact maps were unable to assign to all the nonnatives a higher energy than that of the native. The conformational model and potentials used by them are all-or-none, while those used here are distance-dependent. Although differences in the two models preclude a direct comparison, we suspect the all-or-none contacts contributed to their results.

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. It seems the current potential can be easily trained better (i.e., so that the  $minE$  native = the native) by adding some constraint inequalities, without changing the general property of the current potential or causing infeasibility of the system. It may be not so difficult to train for a couple of extra proteins, too.

## 4.3. Drawbacks and conclusion

There are some limitations or inconveniences to this approach. Some contact energies in Fig. 2, 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 any linear combination of those has positive values at short distances and gradually reaches zero at long distances. This condition would be fairly easily realized if a large number of basis functions are employed, only with an even larger number of adjustable parameters,  $\{x_{i,t_j}\}$ , 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.

Also, because of the nature of the optimization method, some  $e_{i,t_j}(r_{ij})$  are undetermined if the atom pairs of  $t_i$  and  $t_j$  do not exist in the native. The parameters, therefore, should be adjusted for a couple of proteins to be more generic. That may, however, result in an infeasible set of inequalities, requiring revisions in atom classifications or the basis functions.

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  $\rho$  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  $\Delta E$ s of those natives are small positive values and satisfy the condition on  $\Delta E$  automatically, there still is a possibility that some natives have large, positive (or even negative)  $\Delta E$ . We are testing an additional condition that  $\Delta E$ s among natives be minimized.

In summary, we have shown a new way to obtain realistic, yet simple protein folding potentials 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 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. Simulations using this framework will provide interesting information on protein folding.

## 5. Acknowledgements

This work is supported by NSF (DBI-9614074) and NIH (GM59097-01), and is in debt of all the PDB contributors and administrators. The authors are grateful to Dr. V. N. Mayorov, as some programs used here are the revised versions of his originals. Anonymous referee's comments are also highly appreciated.

## 6. References

- [1] Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (1993). *Nonlinear Programming: Theory and Algorithms*, 2nd ed., John Wiley & Sons, Inc., New York.
- [2] Bernstein, F. C., Koetzle, T. F., Williams, G. G. B., Meyer, F. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535-542.
- [3] Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.
- [4] Bryant, S. H. & Lawrence, C. E., (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**, 92-112.
- [5] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein-folding: a synthesis. *Proteins*, **21**, 167-195.
- [6] Crawford, O. H. (1999). A fast, stochastic threading algorithm for proteins. *Bioinformatics*, **15**, 66-71.
- [7] Crippen, G. M. (1996). Easily searched protein folding potentials. *J. Mol. Biol.*, **260**, 467-475.

- [8] Deutsch, J. M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.*, **76**, 323-326.
- [9] Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, **12**, 183-210.
- [10] Godzik, A., Kolinski, A., & Skolnick, J. (1995). Are proteins ideal mixtures of amino-acids? Analysis of energy parameter sets. *Protein Sci.*, **4**, 2107-2117.
- [11] Hao, M. & Scheraga, H. A. (1996). Optimizing potential functions for protein folding. *J. Phys. Chem.*, **100**, 14540-14548.
- [12] Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., & 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.
- [13] Hestenes, M. R. (1980). *Conjugate direction methods in optimization*, Springer-Verlag, New York.
- [14] Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522-524.
- [15] Huang, E. S., Subbiah, S., & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, **252**, 709-720.
- [16] Huang, E. S., Subbiah, S., Tsai, J. & 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.
- [17] Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A34**, 827-828.
- [18] Klimov, D. K. & Thirumalai, D. (1998). Cooperativity in protein folding: from lattice models with sidechains to real proteins. *Fold. Design*, **3**, 127-139.
- [19] Lathrop, R. H. & Smith, T. F. (1996). Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.*, **255**, 641-665.
- [20] Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, **227**, 876-888.
- [21] Maiorov, V. N. & Crippen, G. M. (1995). Size independent comparison of protein three-dimensional structures. *Proteins*, **22**, 273-283.
- [22] Mirny, L. A. & Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.*, **264**, 1164-1179.
- [23] Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534-552.
- [24] Miyazawa, S. & 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.
- [25] Momany, F. A., McGuire, R. F., Burgess, A. W., & Scheraga, H. A. (1975). The mean geometry of the peptide unit from crystal structure data. *Biochim Biophys. Acta*, **359**, 298-302.
- [26] Park, B. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, **249**, 493-507.
- [27] Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, **258**, 367-392.
- [28] Park, B., Huang, E. S., & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, **266**, 831-846.
- [29] Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, Springer-Verlag, New York.
- [30] Shortle, D., Simons, K. T., & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small protein. *Proc. Natl. Acad. Sci. USA.*, **95**, 1158-1162.
- [31] Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.*, **257**, 453-469.
- [32] Torda, A. E., (1997). Perspectives in protein-fold recognition., *Curr. Opin. Struct. Biol.*, **7**, 200-205.
- [33] Vajda, S., Sippl, M., & Novotny, J., (1997). Perspectives in protein-fold recognition., *Curr. Opin. Struct. Biol.*, **7**, 222-228.
- [34] Vendruscolo, M. & Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, **109**, 11101-11108.
- [35] Wang, Y., Zhang, H., Li, W., & Scott, R. A. (1995). Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl. Acad. Sci. USA*, **92**, 709-713.
- [36] Yue, K. & Dill, K. A. (1995). Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA.*, **92**, 146-150.