PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys

Mehdi Mirzaie,[1] Changiz Eslahchi,[1]* Hamid Pezeshk,[3] and Mehdi Sadeghi[2,4]

[1] Department of Mathematical Sciences, Shahid Beheshti University, Post Code 1983963113, Tehran, Iran
[2] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[3] School of Mathematics, Statistics and Computer Science, College of Science, Center of Excellence in Biomathematics, University of Tehran, Tehran, Iran
[4] National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

ABSTRACT

The purpose of this article is to introduce a novel model for discriminating correctly folded proteins from well designed decoy structures using mechanical interatomic forces. In our model, we consider a protein as a collection of springs and the force imposed to each atom is calculated. A potential function is obtained from statistical contact preferences within known protein structures. Combining this function with the spring equation, the interatomic forces are calculated. Finally, we consider a structure and define a score function on the 3D structure of a protein. We compare the force imposed to each atom of a protein with the corresponding atom in the other structures. We then assign larger scores to those atoms with lower forces. The total score is the sum of partial scores of atoms. The optimal structure is assumed to be the one with the highest score in the data set. To evaluate the performance of our model, we apply it on several decoy sets.

Proteins 2009; 00:000–000.

Key words: knowledge-based potential; interatomic force; score function; spring.

INTRODUCTION

The current approaches to protein structure prediction are based on the thermodynamic hypothesis according to which native state of protein is at the lowest free energy state under physiological condition. Thus, using energy function to detect a correct protein fold from incorrect ones is very important for protein structure prediction and protein folding. Commonly, two different types of potential energy functions have been used either for the identification of native protein models from a large set of decoys, or protein fold recognition and threading studies.[1–10] The first types of potentials are based on the fundamental analysis of the forces between the particles referred to as physical energy function. The second types are knowledge-based energy function and are based on information from known protein structures. In physical energy function, a molecular mechanics force field is used. Molecular mechanics force fields are parameterized from ab initio calculation and small molecular structural data. They are essentially summation of pairwise electrostatic and Van der Waals interaction energies, bonds, angles and dihedral angles terms.[11–14] In addition, terms such as entropy and solvent effect are implicitly included. Although physical energy function is widely used in molecular dynamic simulation of proteins, these functions have been out of favor in protein structure prediction because of their greater computational costs. To reduce computational complexity of the protein folding problem, knowledge-based or empirical mean force potential is widely used. The structure of folded proteins reflects the energy of the interaction of all their components, including all enthalpic and entropic contributions, as well as solvent effects.

Such potentials provide an excellent shortcut toward a powerful objective function. It can be used to coarse grain the system to obtain potential between groups of atoms by the use of experimentally determined structures. In this approach, statistical thermodynamics is used in an analysis of the frequency of observed state in order to approximate the underlying free energy.[15]

Most often, the distribution of pairwise distances are used to extract a set of effective potential between residues or atoms. This distribution can be complied from the protein structure database and by defining a reference state. Boltzman's law is used to calculate the interaction energy of a particular pair. The total potential energy of a protein is then simply taken as a sum over all pairwise interactions.

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: IPM; Grant number: CS 1385-1-02.
*Correspondence to: Changiz Eslahchi, Department of Mathematical Sciences, Shahid Beheshti University, Post Code 1983963113, Tehran, Iran. E-mail: ch-eslahchi@sbu.ac.ir.
Received 23 October 2008; Revised 12 March 2009; Accepted 19 March 2009
Published online 20 April 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22457

In most cases, one or two points for each residue are considered to represent a protein.[16–18] These points are usually C_α and C_β or the center of mass of each side chain. The interaction can be either distance-dependent or only dependent on contact. A large variety of knowledge-based potential of mean force has been developed by introducing additional interactions such as surface area terms, probability of main chain and side chain dihedral angles and heavy atoms.[19–35]

Protein potentials based on Delauney tessellations were originally proposed by Tropsha and coworkers.[36,37] The amino acids in a protein chain are represented by their C_α or C_β or all atoms. Then the Delauney tessellation of the resulting point set is computed using Qhull.[38] It defines nearest neighbors in an unambiguous and parameter free way by decomposing the structure into tightly packed tetrahedra which have the atoms as their corners. From the database of residue frequencies for these tetrahedra, a potential function is extracted using the *inverse Boltzmann law.*[15]

In this study, we introduce a new method based on stability of protein structure and use our idea in fold recognition. We consider a protein as a system of atoms and the force that atom i imposes to atom j is calculated by the Hooke's law. That is, $F = -Cx$, where $C > 0$ is the spring constant, x is equal to $|x_{\text{rest}} - d|$, where x_{rest} is distance between two atoms i and j in equilibrium (distance at which the energy between two atoms is minimum, that is, the force between two atoms is zero), and d is the distance between two atoms i and j. The force constant C is not known, but we can calculate E_d^{ij}; mean force potential for atoms i and j at distance d. So by combining $F = -Cx$ and $E^{ij}_{d_{E^{ij}_d}} = C\frac{x^2}{2}$ we have the absolute value of the force, $F = 2\frac{E^{ij}_d}{x}$. Direction of the force, discussed in materials and methods, is also considered. Therefore by considering an ideal spring we can calculate resultant of forces imposed to each atom. In a native structure, we expect the force on each atom to be close to zero.

We begin by deriving a mean force potential from a sample of the native structures. Next, we derive the force that atom i imposes to atom j, and then we introduce a scoring function. Our results based on the energy and force are compared in results and discussions. At the end, we compare our theory to six other scoring functions with the aid of six multiple target decoy sets. To test the performance of our approach we apply it on several decoy sets to measure its ability to discriminate native structure from decoys. Several decoy sets, which contain hundreds of decoy proteins and generated in different ways have been used. In most cases this approach has been able to distinguish native structure from decoys. The calculated Z-score, as a useful quantitative measure of the validity of the calculated potential, shows high value for all protein data sets.

A detailed description of the training and decoy sets, the scoring functions and the evaluation criteria are presented in results and discussion.

MATERIALS AND METHODS

Training and decoy sets

Training set is TOP500H that contains 500 nonredundant proteins with resolution of at least 1.8 Å.[39] In this data set each two proteins have at most 60% sequence identity. Proteins presented within any of publicly available decoy sets have been used to test our method. The test set containing the *ig_structal_hires, 4state_reduced, fisa_casp3, fisa, vhp_mcmd, semfold, hg_structal, lmds, ig_structal and lattice_ssfit* obtained from Decoys'R'us (http://dd.compbio.washington.edu).

Atom type definition

The first step in our approach is to define the different atom types for all the heavy atoms of the 20 amino acids. In a strict physico–chemical point of view, all the atoms with different environments, connectivity and chemical nature, would be different. Among all the 20 amino acids the total number of heavy atoms is 167 and the number of nonequivalent heavy atoms is 98. To reduce this number and raise the observed frequencies, we have considered four atom types containing C (carbon), N (nitrogen), O (oxygen), and S (sulfur).

Calculation of distance-dependent potential

We extract knowledge-based mean force potential for assessment of contact of four atom types within TOP500H records. To quantify contact preferences, we use Delaunay tessellation. In fact, Voronoi tessellation partitions the space into convex polytopes called Voronoi polyhedra. For a given protein the Voronoi polyhedra is the region of space around an atom, such that all points of this region are closer to this atom than to any other atoms of the protein. A group of four atoms, whose Voronoi polyhedra meet at one vertex, forms another basic toplogical object called, the Delaunay tessellation simplex. So we consider two atoms in contact, if they are two vertices of an edge in a simplex.

The distances between any two atoms are divided into 29 distance shells, starting from 0.75 Å with distance shell 0.5 Å in width. All pairwise occurrences out of this range are excluded. Another parameter that we consider is the sequence separation. We consider 25 distinct values for sequence separation as locals ($1 \leq k \leq 24$) and nonlocals ($k \geq 25$). This definition is based on the observation of the energy curve. All pairwise atoms of N and C with a sequence separation equals one are also omitted. We count each pairwise contact asymmetrically, which means that a contact between atom A and atom B (when A is closer to the N-terminal side than B) is counted separately from a contact between atom B and atom A (when B is closer to the N-terminal side than A). For that reason, we have constructed 400 different atom pairwise empirical distributions. It should be noted here that to

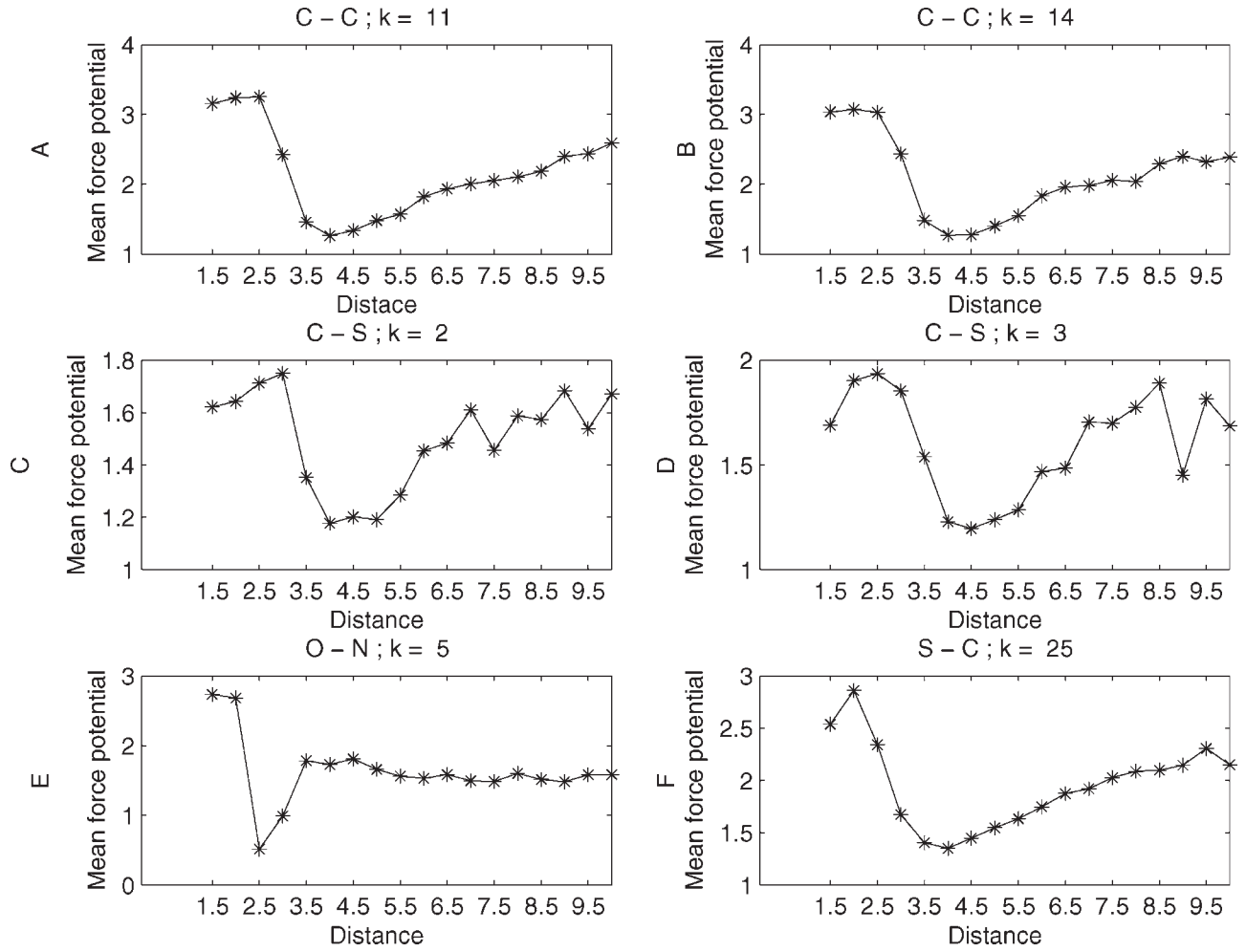

Figure 1

The energy function $\Delta E_k^{ij}(l)$: (**A**) for the atomic pair C—C at sequence separations of k = 11; (**B**) for the atomic pair C—C at sequence separation of k = 14; (**C**) for the atomic pair C—S at sequence separation of k = 2; (**D**) for the atomic pair C—S at sequence separation of k = 3; (E) for the atomic pair O—N at sequence separation of k = 5; (F) for the atomic pair S—C at sequence separation of k = 25.

avoid obtaining zero frequencies for pairwise atoms, we have added one count to all frequencies. The calculation of pairwise pseudo-energy terms has been carried out as described by Sippl.[15] Therefore we use following expression for pseudo-energy for atomic pair i and j at sequence separation k and at distance shell l:

$$\Delta E_k^{ij}(l) = RT\ln[1+M_{ijk}\sigma] - RT\ln\left[1+M_{ijk}\sigma\frac{f_k^{ij}(l)}{f_k^{xx}(l)}\right], \quad (1)$$

where M_{ijk} is the number of observations for the atomic pair i and j at sequence separation k and is equal to:

$$M_{ijk} = \sum_{l=1}^{29} f(i,j,k,l),$$

σ is the weight given to each observation. We considered σ = 1/50,[15] that means on 50 observations $f_k^{ij}(l)$ and $f_k^{xx}(l)$ have the same weight for the calculation of $\Delta E_k^{ij}(l)$. It has been proposed that lower values for σ might perform better.[35] In our case, due to the higher number of observations for the different atom pairwise contacts, this parameter has little effect on the results.

$f_k^{ij}(l)$ is the relative frequency of occurrence for the atomic pair i and j at sequence separation k in the class of distance l and is equal to:

$$f_k^{ij}(l) = \frac{f(i,j,k,l)}{M_{ijk}}$$

$f_k^{xx}(l)$ is the relative frequency of occurrence for all the atomic pairs at sequence separation k in the distance shell l and can be expressed by:

$$f_k^{xx}(l) = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} f(i,j,k,l)}{\sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{l=1}^{29} f(i,j,k,l)}$$

The temperature is set to 293 K, so that RT is equivalent to 0.582 kcal/mol.

Calculation of forces imposed to an atom

Figure 1 shows the energy curve for some pairs of atoms at some particular sequence separation. The shape of energy curve justifies our model. As shown by Figure 1, energy curve around minimum is U shaped. So we can assume energy function around minimum follows Hook's law. From the laws of mechanics, it is well known that the force on the end of a spring at rest is zero.

In fact at $x = 0$ the spring is at equilibrium and the force on the end of the spring is zero. In other cases, the quantity of the force at x_0 is obtained from

$$F = -Cx_0, \tag{2}$$

where C is the spring constant and x_0 is the displacement from $x = 0$ (equilibrium state) to $x = x_0$.

The energy needed to move the end of spring from $x = 0$ to $x = x_0$ is equal to

$$E = \int_0^{x_0} Cx\mathrm{d}x = C\frac{x_0^2}{2} \tag{3}$$

So by (2) and (3) we have

$$F = -\frac{2E}{x_0} \tag{4}$$

The absolute value of F is

$$F = \frac{2E}{|x_0|} \tag{5}$$

The force represents the rate of change of energy. So there is a tendency to move to minimum energy state. The energy is minimized whenever the force is zero. In that case the spring is said to be at equilibrium. In our model, each pairs of contacted atoms are considered as ends of a spring and the quantity of the force imposed to each atom is calculated using (5) with respect to types of atoms and the distances among them.

Let E_0 be $\Delta E_k^{A_iA_j}(l)$ that obtained from (1) and e_m be the global minimum of E_0 in the range of distances between 0.75 to 15 Å. Let b_m be the bin of distance in which E_0 is minimum. We consider $E = E_0 - e_m$ and at b_m, we have $E = 0$.

Thus the quantity of the force that atom A_i imposes to atom A_j is

$$F = \frac{2E}{|d - d_0|},$$

where in this case $d_0 = b_m$.

If $d_0 < d$, the direction of the force is from A_i to A_j and if $d < d_0$, the direction of the force is from A_j to A_i. If $d = d_0$ then the force is zero.

Given the coordinates of A_i and A_j, the quantity of the force and the direction of it, we can calculate vector of force. For instance, let the direction of the force which is imposed to A_j be from A_i to A_j and the quantity of it be F_{ij}. The unit vector from A_i to A_j, u, is

$$u = \frac{((x_j - x_i), (y_j - y_i), (z_j - z_i))}{\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}}$$

So, the vector of force is F_{ij} u.

Now, for the calculation of the force imposed to atom A_i, we first determine neighborhood of A_i by using Delaunay algorithm and then calculate the vectors of forces which imposed to A_i from each of the neighbors and consider the resultant of these as the force imposed to A_i. The energy of each pair of atom types in different sequence separation and distance interval are given in Supporting Information S1.

RESULTS AND DISCUSSION

The data set of proteins used for the extraction of the statistical potential is Top500H. This is a nonredundant set of 500 proteins resolved by X-ray crystallography with at least 1.8 Å. Contacts are quantified by using a constrained Delaunay tessellation procedure. After performing calculations, 4,127,409 atomic pairwise contacts have been obtained. The lowest occurrence of atomic pairwise contacts are between two S's with 882 observations and the highest occurrence of atomic pairwise contacts are between two C's with 1229293 observations.

The total number of observations is strongly dependent on the sequence separation. Figure 2 shows the average of total observations for all the different sequence separations. The total number of atom-atom pairwise observations, considering the entire atom types in all distance shells at a particular sequence separation, has been

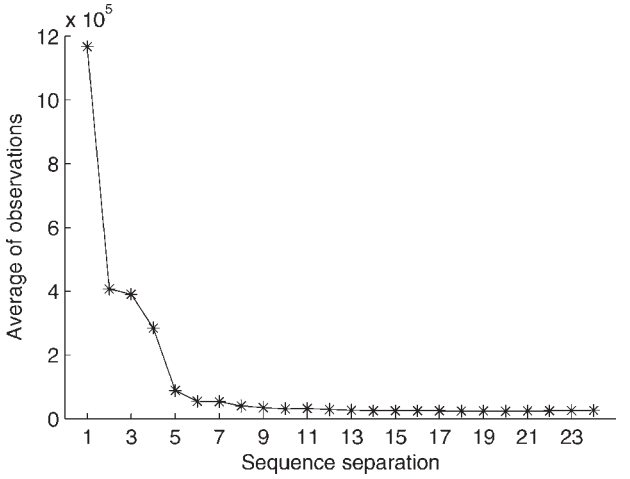

Figure 2
The average frequency of observations versus sequence separation.

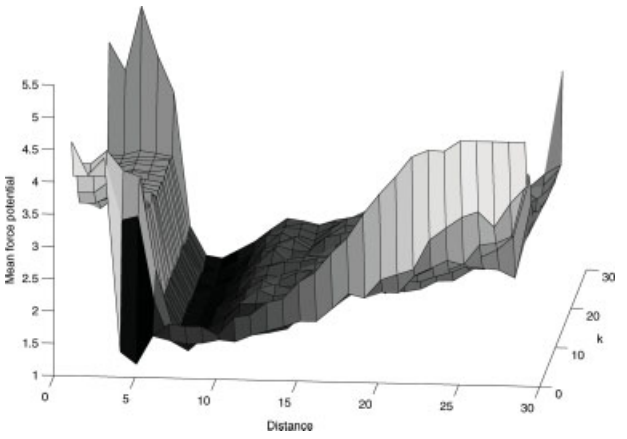

Figure 3

Plot of the energy $\Delta E_k^{ij}(l)$ versus the distance and the sequence separation for two C's.

divided by 16. Figure 3 shows the energy curve of this potential for two C's in the whole range of sequence separation. The atomic pairwise energy shows a well defined function with a deep minimum.

Atoms of adjacent residues that are covalently connected are close in space, therefore sequentially local interactions should be separated from sequentially nonlocal interactions when pairwise potential is calculated.[40] The shapes of pairwise energy curves for short sequence separations exhibit variability and for distant sequence separations monotonic curves are observed. The best discrimination between natives and decoys is achieved when we derive and test locals ($1 \leq k \leq 24$) and nonlocals ($k \geq 25$) potentials and forces together. Usually, sequence separations longer than about 10 amino acids are considered as nonlocal and not split.[41] In our work, we need to consider exact minimum energy of each pair of atom types and then based on energy curves, sequence separations up to 24 residues are taken separately.

Energy and force calculation

For a given 3D structures of a protein, the total potential of mean force, G, is

$$G = \frac{1}{2}\sum_{i,j}\Delta E_k^{ij}(l),$$

where the summation is over all pairs of atoms. For calculation of the force imposed to each atom we determine neighboring atoms of any atom by Delaunay tessellation procedure. For every pair of atoms at any sequence separation we consider an interval with radius 2.5 Å around minimum as a valid region and then we exclude neighbor atoms with distances out of the valid region. Finally we calculate the force imposed to each atom with remaining atoms.

Calculation of pairwise atom potential and force for proteins in decoy sets

For each structure in the decoy sets, including native state and decoys the total potential G is calculated. The structure with lowest energy is considered as native structure. Native state is correctly identified if its structure has the lowest value of G. To identify of a native structure from decoys based on forces imposed to each atom of proteins we use a scoring function.

Let S be a system of particles that is composed of n particles $1, 2, \ldots, n$ and $F_1, F_2, \ldots, F_n$, be n force vectors on each particle. We know from mechanics that when a system is at equilibrium, the force on each particle is zero. So if we think a protein as a system of atoms in equilibrium or near equilibrium and calculate the force on each atom, we expect in a strict physical law that the force on each atom in native protein should be zero or close to zero. But this assumption about protein is unrealistic and we need a similar definition of equilibrium. First we consider a protein P as a system composed of n particles (n is the number of atoms of P). Then we expect that S is a good approximation of the native structure of a protein if the force on each atom of protein with this structure is near zero. For this purpose, we consider the following procedure:

1. Standard forces imposed to each atom in each protein in decoy data sets are calculated.
2. Let n be the number of decoys of protein and m be the number of atoms of protein. For each atom, for example, i'th atom, we find the minimum of force on this atom among n proteins. Let m_i be this minimum.
3. Let for k number of proteins, the force on i'th atom be equal to m_i. We add the score $1/k$ to each of these proteins.
   We repeat steps 2 and 3 for every m atom to calculate the score of a protein.
4. We calculate the median of scores and remove proteins with scores lower than median. Now, we have a new set of proteins.
5. We repeat Steps 2 and 3 twice.
   We expect to find the native structure with the highest score.

Performance of pairwise atom potential and force

Five performance measures are considered to evaluate the performance of the model.

1. *Rank native,* we rank proteins according to their scores. The score of native structure should be maximized among decoys in force model and minimum in energy model. Ideally the rank of the native structure should be 1.

Table I
Performance of Energy and Force Models on Decoy Sets

Decoy Source	Average no. of decoys per target	Energy model			Force model		
		Top 1	Top 2	Top 5	Top 1	Top 2	Top 5
4state	666	3/7 (2.81)	5/7	5/7	1/7 (2.11)	2/7	3/7
lattice-ssfit	2000	8/8 (4.8)	8/8	8/8	8/8 (11.26)	8/8	8/8
lmds	435	4/10 (1.56)	4/10	4/10	10/10 (21.93)	10/10	10/10
fisa	500	3/4 (5.27)	3/4	3/4	4/4 (21.52)	4/4	4/4
fisa_casp3	1432	3/6 (3.55)	5/6	5/6	3/6 (5.75)	3/6	3/6
hg_structal	30	20/29 (2.17)	22/29	23/29	11/29 (0.52)	18/29	24/29
ig_structal	61	21/61 (1.12)	24/61	27/61	13/61 (1.45)	35/61	50/61
ig_structal_hires	20	14/20 (2.39)	14/20	16/20	10/20 (1.52)	12/20	17/20
vhp_mcmd	6256	0/1 (−1.9)	0/1	0/1	0/1 (0.3)	0/1	0/1
semfold	13,038	5/6 (2.76)	5/6	5/6	4/6 (10.6)	5/6	5/6
Total	2443	81/152 (2.45 ± 2.01)	90/152	96/152	64/152 (7.69 ± 8.38)	97/152	124/152

The numbers in parentheses are the average Z-scores.

2. RMSD of the best scoring conformation. Since decoys are well constructed and there are native like conformations in these sets, so the RMSD can be used to assess our model in recognition of native-like conformation.
3. Correlation Coefficient (*CC*) between score and RMSD. The percentage of detected native state is a necessary criterion but not sufficient. Because it does not describe the correlation between the score of a model and its structural similarity to the native state, suggested by the funnel-shaped free energy landscape of protein folding. We expect that in the force model, proteins with high RMSD have low scores; that is, *CC* should be close to −1 and in the energy model proteins with high RMSD have high scores; that is, *CC* should be close to 1.
4. *Z-score*. The *Z-score* of the native structure in the decoys set is equal to

$$Z\text{-score} = \frac{<\text{score}_{\text{decoys}}> - \text{score}_{\text{native}}}{\sigma_{\text{decoys}}},$$

in which $\text{score}_{\text{native}}$ is the score calculated for native structure and $<\text{score}_{\text{decoys}}>$ and σ_{decoys} are the average and the standard deviation of scores distributions of decoys proteins, respectively.
5. Fraction Enrichment (*FE*). The Fraction Enrichment is another measure that is the percentage of the top 10% lowest RMSD structures which are found also in the top 10% best scoring ones. Identification of the nonnative structure closest to the native structure among a set of decoys is more difficult than identification of the native structure. This measure should be close to 1 when a sufficient number of near native decoys has high score.

We test our approach on decoy sets available in the Decoys'R'us database under the category 'multiple'. These decoys are made with different methods and are very appropriate for assessment of model. The performances of the energy and force model are compared in Table I. The results are presented in terms of the average Z-score and the number of first ranked (Top 1), second ranked (Top 2) and fifth ranked (Top 5) native structures within the decoy sets. Although energy model has distinguished the higher number of native structures as Top 1, but the force model has the higher average Z-score.

Also, if a success is defined by ranking the native structure as one of the two or five highest score (Top 2 or Top 5), the performance of force model is remarkably better than energy model. The details of results obtained by energy model and force model on each protein in the decoy sets are summarized in Supporting Information S2.

For data sets where the whole range of RMSD is reported, the correlation coefficient should be positive for energy model and negative for the force model. Also it needs to be significantly different from 0 and similarly the fraction enrichment (FE) should be significantly larger than 10%.

The set *4state-reduced* contains 7 different proteins. For each protein, 632 to 689 decoy conformations are presented in the data set. This decoy set has been generated using a four-state off lattice model with a conformational relaxation method.[42] The energy model does particularly well with *4state-reduced*. Three of the native structures could be recognized and the *Z-score* is high. The average *CC* is 0.41 and *FE* is larger than 10%, but the force model is not as good as the energy model, although *Z-scores* are low and *FE*s are larger than 40 percentage.

The *fisa* and *fisa_casp3* decoy sets with four and six targets (500 and 1400 decoy per target), respectively, have been obtained using a combination of a Bayesian scoring function and a simulated annealing protocol.[43,44] For *fisa* data set the scores in the force model are better than energy model. For *fisa-casp3* the performance of the energy model is the same as the force model but corresponding *Z-scores* are very low.

Table II
Assessments of Different Potential Functions by the Rank Native Structure in Five Decoy Sets

Source	DFIRE	Rosetta	ModPipe-Pair	ModPipe-Surf	ModPipe-Comb	DOPE	PC2CA	Force model	Energy model
4state-reduced									
1ctf	1	1	1	1	1	1	1	1	1
1r69	1	2	1	17	1	1	1	8	1
1sn3	1	1	1	7	1	1	1	23	7
2cro	1	5	1	103	1	1	1	4	1
3icb	4	6	15	33	8	1	1	2	2
4pti	1	1	1	71	1	1	1	13	47
4rxn	1	1	1	18	1	1	667	85	2
Fisa									
1fc2	254	158	491	1	453	357	1	1	28
1hdd-C	1	90	293	18	135	1	1	1	1
2cro	1	26	11	146	19	1	1	1	1
4icb	1	1	196	2	167	1	1	1	1
Fisa-casp3									
1bg8-A	1	1068	1	1180	282	1	1	244	2
1bl0	1	960	4	912	86	1	1	—	1
1jwe	1	1177	1	1119	6	1	1	1	1
Lmds									
1b0n-B	430	300	56	186	18	34	1	1	19
1bba	501	174	501	117	444	501	501	1	44
1fc2	501	291	325	54	222	476	53	1	1
1ctf	1	1	1	1	1	1	1	1	501
1dtk	1	9	4	1	1	1	2	1	7
1igd	1	1	1	3	1	1	1	1	332
1shf-A	1	5	24	18	7	1	1	1	1
2cro	1	2	4	28	12	1	1	1	1
2ovo	1	29	5	8	2	1	1	1	1
4pti	1	4	1	44	1	1	1	1	30
Lattice-ssfit									
1beo	1	1	1	1	1	1	1	1	1
1ctf	1	1	1	1	1	1	1	1	1
1dtk-A	1	1	1	35	1	1	1	1	1
1fca	1	1	1	4	1	1	1	1	1
1nkl	1	1	1	1	1	1	1	1	1
1pgb	1	1	1	3	1	1	1	1	1
1trl-A	1	45	1	123	1	1	1	1	1
4icb	1	1	1	3	1	1	1	1	1
Total correct prediction	27	14	19	7	18	28	27	24	20

The data sets *ig_structal, hg_structal* and *ig_structal_-hires* contain immunoglobulins (ig) or globins (hg) have been created by homology modeling. The energy model perform better than force model, but force model performance is better in finding native structures as Top 5, which are most native-likes. For these data sets, rank native, *Z-scores*, *CC* and *FE* justify energy and force models well.

The largest *lattice_ssfit* decoy set, containing 2000 decoys for each of the eight targets, has been generated using a tetrahedral lattice model with the all-atom ENCAD energy function.[45] The ranges of RMSD for all eight proteins in data set are larger than 4 Å. The results of force model for this group of decoys are more acceptable than energy model. Although the force and the energy model identify all of the eight structures, the *Z-scores*, *CC* and *FE* for force model are more acceptable than the energy model.

The *lmds* set includes decoys with RMSD's less than 10 Å. The *lmds* decoy set with 215–500 models for each of 10 primarily short targets, has been obtained by a local optimization method and a reduced ENCAD energy function.[46] The force model does particularly well with this group of decoys. All of the five criteria justify force model.

The *semfold* set include a very large number of decoys for each of the six proteins. In some cases RMSD from native are in range 3 Å to 5 Å. This decoy set has been generated by fragment insertion method. This decoy set is the most challenging decoy set, since it has more than 10,000 decoys for each of the six targets. The energy model and the force model could recognize five and four native structures, respectively.

The *vhp_mcmd* decoy set has been generated by molecular dynamics simulations. For this set the force model is better than the energy model, but *FE*'s and *CC*'s for energy model are plausible. All the target native structures, in all decoy sets have been determined by X-ray crystallography, except for 1bba in *lmds* which has been determined by NMR spectroscopy.

Table III
Performances of Energy and Force Models on Decoy Sets When Native Structures are Omitted

Decoy Source	No. of decoys per target	RMSD range	Average	Std Dev	Energy model		Force model	
					Top 1 RMSD	Z-score	Top 1 RMSD	Z-score
4state								
1ctf	630	1.3–9.1	5.1	1.7	3.8	0.76	2.8	1.3
1r69	676	0.9–8	4.8	1.5	3.1	1.13	4.3	0.33
1sn3	660	1.3–9.1	5.7	1.7	4.6	0.64	2.4	1.94
2cro	673	0.8–8.3	4.6	1.6	1.9	1.69	0.9	2.3
3icb	654	0.9–9.4	5.4	2	1.9	1.75	1.7	1.85
4pti	686	1.4–9.2	5.5	1.6	2.6	1.81	2.6	1.181
4rxn	677	1.4–8.1	5.3	1.6	4.6	0.44	2.4	1.81
Lmds								
1bon-B	498	2.4–4.9	3.7	0.4	3.6	0.25	3.9	−0.5
4pti	344	4.8–11.8	7.8	1.4	6.7	0.79	10.6	−2
1ctf	496	3.6–11.7	7.9	2	10.2	−1.15	8.4	−0.25
1bba	501	2.8–7.5	4.7	0.9	3.1	1.77	4.6	0.11
1dtk	216	4.3–11.9	8.2	1.4	8	0.14	7.9	0.21
1fc2	501	4–8.4	5.5	0.9	6	−0.55	4.3	1.33
1shf	437	4.4–11.2	8.5	1.1	9.1	−0.55	9.4	−0.81
2cro	501	3.9–11.4	8.7	1.5	10.4	−1.13	10.5	−1.2
1igd	501	3.1–12.1	7.6	1.6	6.7	0.56	6.1	0.93
2ovo	348	4.4–12.5	9.1	1.4	8.7	0.29	10.7	−1.14
vhp_mcmd								
1vii	6256	0.5–12.4	6.3	2	6	0.15	0.85	2.72
semfold								
1ctf	11,402	0.1–12.5	9.4	1.3	0.1	7.15	1.1	6.46
1eh2	11,442	0.3–14.9	10.6	1.5	0.9	6.46	1.3	6.23
1pgb	11,282	0.1–12	9.3	1	0.1	9.2	0.1	9.2
1nkl	11,662	0.2–13.9	9.3	1.6	9.9	−0.38	10	−0.43
1khm	21,081	3.8–14.5	9.8	1.8	7.1	1.5	8.2	0.89
1e68	11,362	0.1–11.8	8	1.6	0.9	4.31	0.7	4.56

There are some reasons for discrimination of native structure using knowledge-based potential functions, as discussed by Shen and Sali.[34] We compared our approach based on energy and force to seven previously published scoring functions, including DFIRE,[33] Rosetta,[47–49] ModPipe-Pair,[50] Modpipe-surf,[50] DOPE[34] and PC2CA.[32] Table II compares the performances of different methods together with our energy and force models in recognizing native structures from decoys in some decoy data sets. Our approach based on force does not work well with the *4-state reduced*, but does well with *lmds*. None of the other previously published scoring functions based on energy in Table II and our approach based on energy work well on *lmds*, but force score can discriminate all of the 10 native structures. It is noticeable that only our force model together with PC2CA and Modpipe-surf has been able to identify 1fc2 in the *fisa* set. Our approach based on force correctly identifies 24 native structures for 32 target in five multiple target decoy sets, while the approach based on the energy identifies 20 native structures.

To evaluate the performance of energy and force models when performing ab initio folding, one experiment is made to discriminate between the near native and nonnative structures when the native structure is omitted from decoy set. Since different data sets have different ranges of RMSD for nonnative proteins, we calculate Z-score for the structure that is distinguished as Top 1. Table III shows the results of this test for four decoy sets containing more decoy structures. Average and standard deviation of RMSD of decoy structures and the RMSD of structures detected as Top 1 and also Z-score for each protein in data sets are shown in Table III. Positive Z-scores show that regarding different programs to generate decoys, energy and force model have ability to detect near native structures among all decoys although they are not the best structures with lowest RMSD. Due to the sensitivity of the force model to perturbations from native protein and also our method to rank proteins based on scoring function for forces imposed to each atom, we can not detect near native structures (low RMSD) based on force model in the data sets in which decoys are far from native structures. Therefore the results of energy model are better than force model but in the presence of very native like structures in the data set, the force model performs better.

Table IV
The Averages and the Variances of the Forces on the Atoms in Three Secondary Structures; Alpha, Beta, and Coil

Secondary structure	Average	Variance
Alpha	1.5761	1.2851
Beta	1.6561	1.2978
Coil	1.6543	1.432

Kruskal Wallis test for comparing forces imposed to the atoms of alpha, beta, and coil

It is well known that long loops tend to be more flexible than regular secondary structures such as helices and strands. Using Debye-Waller factor (B-values) as measures for local residue flexibility, it has been shown that residues in regular secondary structure (helix and strand) tend to occupy regions of lower B-values, while residues in nonregular secondary structure occupy regions with higher B-values.[51] Based on force model, it is expected that the force imposed to atoms in the secondary structures elements of alpha helices and beta strands, should be different from atoms in coil. Even, the force imposed to atoms in alpha helices must be different from atoms in beta strands. To examine the hypothesis of variation among these three groups we use the Kruskal Wallis one way analysis of variance. This is a nonparametric method based on statistical analysis for testing equality of the mean of several groups. The results of Kruskal Wallis test for our data set show the significant differences among groups (P-value is < 0.01). Pairwise comparisons of all groups are also performed. It is concluded that all groups are significantly different (P-values are < 0.01). The averages and the variances of the forces on three structures alpha, beta, and coil are shown in Table IV.

The standard force imposed to each atom is calculated using:

$$F_s(a,s) = \frac{F(a,s) - \mu_s}{\sigma_s},$$

where $F(a,s)$ is the force impose to an atom a in secondary structure s, and μ_s and σ_s are the average and the standard deviation of forces in the secondary structure s, respectively.

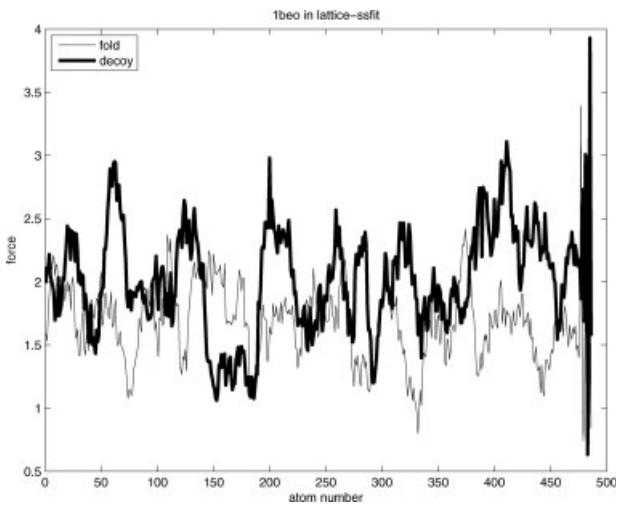

Figure 4

Force profiles of protein (1beo) in their native (continuous line) and misfolded (bold line) conformations. The average of forces imposed to atoms in a window of size 10 is used to calculate force profiles.

CONCLUSIONS

In this study, at first we have constructed a knowledge-based potential function based on four heavy atoms and tested it on decoy data sets to discriminate native structures from decoys. The results show that this potential function has high power to detect the native fold. The goal of this study is not to use the energy for native fold detection but rather, we suppose that the force atom i imposes to atom j can be calculated by considering the atom pair as the ends of spring that obeys Hooke's law. This force is calculated from the mean force potential and finally the force imposes to each atom from its neighboring atoms is calculated. Considering a protein as a system of particles we expect that in native fold, more residues should be at equilibrium or near equilibrium. Figure 4 shows an example of forces imposed to atoms of one protein (1beo) in its native fold and one of its misfolds. As shown by the figure, most of residues undergo less force in native fold. Introducing a scoring function, we evaluated it on several decoy sets to argue its ability to discriminate native folds from decoys. Results show that the method based on interatomic force performs better than energy based methods to detect native or near native folds. This model can be applied in ab initio protein structure prediction, fold assignment, sequence structure alignment and template selection and also in molecular dynamics and normal mode analysis.

ACKNOWLEDGMENTS

The authors would like to thank anonymous referees for their constructive and valuable suggestions.

REFERENCES

1. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr opin struct Biol 1997;7:194–199.
2. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
3. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
4. Hao M, Scheraga H. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
5. Miyazawa S, Jeruigan R. An empirical energy potential with a reference stat for protein fold and sequence recognition. Proteins Struct Funct Genet 1999;36:357–369.
6. Laziridis T, Gill S. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
7. Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized born solvent model. Proteins 2002;48:404–422.

8. Doming BN, Brooks CL. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002;23:147–160.
9. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. Proteins 1999;35:133–152.
10. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
11. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 1983; 4:187–217.
12. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
13. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106:765–784.
14. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
15. Sippl MJ. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
16. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
17. Covell DG. Folding protein α-carbon chains into compact forms by Monte Carlo methods. Proteins: Struct Funct Genet 1992;14:409–420.
18. Sun S. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci 1993;2:762–85.
19. Bauer A, Beyer A. An improved pair potential to recognize native protein folds. Proteins 1994;18:254–261.
20. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
21. Melo F, Feytmans E. Assessing protein structures with nonlocal atomic interaction energy. J Mol Biol 1998;277:1141–1152.
22. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
23. Tobi D, Elber R. Distance dependent, pair potential for protein folding: results from linear optimization. Proteins: Struct Funct Genet 2000;41:40–46.
24. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
25. Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003;329:159–174.
26. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
27. Betancourt MR, Skolnick J. Local propensities and statistical potentials of backbone dihedral angles in proteins. J Mol Biol 2004;342:635–649.
28. Wang K, Fain B, Levitt M, Samudrala R. Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol 2004;4:8.
29. Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. Protein Sci 2005;14:1741–1752.
30. Fang Q, Shortle D. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. Proteins 2005;60:90–96.
31. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
32. Fogolari F, Pieri L, Dovier A, Bortolussi L, Giugliarelli G, Corazza A, Esposito G, Viglino P. Scoring predictive models using a reduced representation of proteins: model and energy definition. BMC Struct Biol 2007;7:15.
33. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue level, pair potential of mean force for folding and binding based on the distance scaled, ideal-gas reference state. Protein Sci 2004;13:400–411.
34. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2407–2524.
35. Bauer A, Beyer A. An improved pair potential to recognise native protein folds. Proteins: Struct Funct Genet 1994;18:254–261.
36. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–221.
37. Munson PJ, Singh RK. Statistical significance of hierarchical multi body potentials based on Delaunay tessellation and their application insequence structure alignment. Protein Sci 1997;6:1467–1481.
38. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Trans Math Software 1996;22:469–483.
39. Lovell S, Davis I, Arnedall W, de Baker P, Word J, Prisant M, Richardson J, Richardson D. Structure validation by C_α geometry: ϕ, ψ, and C_β deviation. Proteins 2003;50:437–450.
40. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
41. Ferrada E, Melo F. Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models. Protein Sci 2007;16:1410–1421.
42. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
43. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
44. Simons KT, Ruczinski I, Kooperberg C, Fox B A, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence- independent features of proteins. Proteins 1999;34:82–95.
45. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
46. Keasar C, Levitt M. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003;329:159–174.
47. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
48. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
49. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci 2006;103:5361–5366.
50. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
51. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005;61:115–126.