HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2024 May 21.

Published in final edited form as:

Proteins. 2011 September ; 79(9): 2648–2661. doi:10.1002/prot.23086.

A statistical mechanics-based method to extract atomic distance-dependent potentials from protein structures

Sheng-You Huang, **Xiaoqin Zou***

Department of Physics and Astronomy, Department of Biochemistry, Dalton Cardiovascular Research Center, and Informatics Institute, University of Missouri, Columbia, MO 65211

Abstract

In the present study we have developed a statistical mechanics-based iterative method to extract statistical atomic interaction potentials from known, non-redundant protein structures. Our method circumvents the long-standing reference state problem in deriving traditional knowledge-based scoring functions, by using rapid iterations through a physical, global convergence function. The rapid convergence of this physics-based method, unlike other parameter optimization methods, warrants the feasibility of deriving distance-dependent, all-atom statistical potentials to keep the scoring accuracy. The derived potentials, referred to as ITScore/Pro, have been validated using three diverse benchmarks: the high-resolution decoy set, the AMBER benchmark decoy set, and the CASP8 decoy set. Significant improvement in performance has been achieved. Finally, comparisons between the potentials of our model and potentials of a knowledge-based scoring function with a randomized reference state have revealed the reason for the better performance of our scoring function, which could provide useful insight into the development of other physical scoring functions. The potentials developed in the present study are generally applicable for structural selection in protein structure prediction.

Keywords

scoring function; protein structure prediction; knowledge-based; distance-dependent all-atom potentials

1 Introduction

Knowledge about the three-dimensional structures of proteins is of high importance to mechanistic studies of protein functions and rational therapeutic design. Currently, the number of the experimentally determined structures in the Protein Date Bank (PDB)[1] is only a small fraction of the amino acid sequences found in the GenBank.[2] Therefore, there is a pressing need to predict protein structures by using computational methods. Over the years, numerous computational approaches have been developed for structure prediction from sequences.[3–9] Roughly, these approaches can be divided into two categories:[10] template-based modeling (comparative modeling and threading), and first-principle modeling (referred as *ab initio* modeling). In template-based modeling, protein

*Corresponding author. zoux@missouri.edu, 573-884-4232 (fax).

structures are constructed based on the known protein structure/fold templates in the PDB that are close to the sequence to be modeled.[11] In first-principle modeling, no homologous or analogous template is needed; instead, three-dimensional protein models are built 'from scratch' by sampling the conformational space and searching for the one with the lowest free energy.[12–15]

One common step for both template-based modeling and first-principle modeling is to first generate a large number of decoy protein conformations. This can be done either by constructing models from different templates or by sampling different regions of the protein conformational space.[16] The availability of an accurate scoring function that can discern the near-native structure from an ensemble of decoy structures is one of the important determinants for the success rate of structure prediction. The number of decoys is huge particularly for first-principle modeling. Thus, the efficiency, in addition to the accuracy, of a scoring function is also important for structure prediction.[17–19]

Despite 30 years of efforts, the scoring function problem remains a great challenge in computational biology.[19] Approaches to the development of scoring functions can be grouped into two broad categories: physics-based and knowledge-based. In physics-based approaches, the energy of a structure is computed as the sum of individual interactions such as van der Waals interactions, electrostatic interactions, and bond stretching, bending and torsional forces, with force field parameters normally derived from quantum mechanical calculations.[20–27] Despite its lucid physical meaning and a number of successes, physics-based scoring functions have not been widely adopted in protein structure prediction due to impractically-high computational cost and insufficient conformational sampling.

An alternative approach is knowledge-based scoring functions, in which empirical energy potentials are derived from the information embedded in the known protein structures.[28–30] Despite their simple forms, the knowledge-based scoring functions appear to be the most successful approaches and are widely used in protein structure prediction.[19,31,32] There are two types of methods to derive potential parameters for knowledge-based scoring functions. The first type of methods is to optimize the potential parameters such that the energy (or Z-score) of the native (or near-native) structure is lower than those of the decoys.[33–41] Despite the success achieved, the parameter optimization approaches may be restricted by two factors. First, some of the potential parameters derived can be unphysical and contradict to chemical knowledge. Second, the high-dimensional optimization process is computationally intensive. To be computationally practical, these methods normally adopt less accurate coarse-grained potentials by using either contact-based (or distance-dependent but with only a few different distances, referred to as quasi-contact-based) potentials or a reduced protein representation.

The second type of methods to derive pairwise potentials in the knowledge-based scoring functions is the potential of mean force (PMF) method. The PMF method directly converts the potentials from the occurrence frequencies of atom pairs in the native structures by using an inverse Boltzmann relation.[42–45] Since the pioneering work of Tanaka and Scheraga,[28] many efforts have been devoted to the use of the PMF method for developing distance-dependent or contact-based potentials at atomic or residue level. The resulting scoring

functions are widely applied to protein structure prediction.[46–50] Despite the success, one unsolved key issue in the PMF method is the determination of the "reference state"[42,43,51] The reference state is defined as the state with no interaction between any two atoms/residues. As pointed out by Thomas and Dill,[42] such an ideal reference state is not achievable due to atom connectivity, excluded volume, and other effects in proteins. Most of the current PMF scoring functions are based on a crude approximation for the reference state by randomly mixing all of the atoms in proteins. Several other useful approximations have also been introduced to characterize the reference state,[48,52] but these are still not sufficiently accurate. Studies on protein-ligand interactions by our group and other groups showed that PMF-based scoring functions can lead to wrong predictions of the native binding modes despite significant success in binding affinity predictions (see refs 53–55 and references therein). It is therefore expected that proper handling of the reference state issue is important to protein structure prediction.

In an elegant work by Thomas and Dill,[44] it was shown that the reference state can be circumvented by an iterative method in a lattice HP model. However, how to circumvent the reference state problem for the much more complicated true protein system remains challenging.

In the present study, we have developed a statistical mechanics-based iterative method to extract distance-dependent, all-atom potentials for structural model selection in protein structure prediction. The method circumvents the long-standing reference state problem by improving the pair potentials iteratively through comparisons of the physics-based pair distribution functions. As shown in the Materials and Methods, the derived potentials rank the native structure with the lowest energy and thus distinguish it from the decoys. Our new scoring function, referred to as ITScore/Pro, has been extensively evaluated with the high-resolution decoy set of 148 proteins by Floudas and colleagues,[36] the AMBER benchmarking decoy set of 47 proteins by Wroblewska and Skolnick,[26] and the CASP8 decoy set of 123 proteins (http://predictioncenter.org/), showing significant improvement in performance. To understand why our scoring function performs well to provide insightful information for development of other physical scoring functions, we have also analyzed how the reference state affects the derived potentials.

2 Materials and Methods

2.1 The iterative method to extract effective potentials

Here, we take a large training set of the experimentally-determined native protein structures and computationally generated ensembles of non-native/decoy structures as the system (see next subsection for details). We use the following paradigm to derive the effective pair potentials of the scoring function by iteration. Similar methods have been used to extract pair potentials from known pair distribution functions for simple liquid systems using an inverse Monte Carlo approach.[56] Figure 1 shows an illustration of our method. The statistical mechanical basis for why the effective potentials derived from the method can discriminate the native structures from the decoys is described in the next section.

The idea of our paradigm is summarized as follows: For a given ensemble of native and pregenerated decoy structures and a set of trial pair potentials $u_{ij}(r)$, we will be able to calculate the pair distribution function $g_{ij}(r)$ for each atom pair ij based on the principles of statistical mechanics (explained below). If $u_{ij}(r)$ are incorrect, the calculated $g_{ij}(r)$ will be different from the true $g^*_{ij}(r)$ observed in the native structures. The corresponding differences $\Delta g_{ij}(r) = g_{ij}(r) - g^*_{ij}(r)$ will be used to optimize $u_{ij}(r)$ by iteration, until $\Delta g_{ij}(r)$ are below a predefined cutoff and thus the trial potentials $u_{ij}(r)$ approach the true effective potentials $u^*_{ij}(r)$. Here, we assume the approximation that all the protein structures (natives and decoys) in the training set form a canonical ensemble. There will be two key issues in this paradigm. First, the initial guess of $u_{ij}(r)$ cannot be far off from the true $u^*_{ij}(r)$; otherwise the iterative process may be trapped at local energy minima. Second, an intelligent iterative function should be proposed for fast convergence of the iterative process.

The procedure for derivation is explained in detail as follows. The fist step is the preparation of the native and decoy protein structures for the training set, which will be described in the next subsection. Notice that decoy generation/preparation is a one-time step.

The second step is the initialization of the potentials, $u^{(0)}_{ij}(r)$, for an iterative process. A good initial guess of the pairwise potentials will make the iterative process efficient by avoiding traps of local minima. In this study, $u^{(0)}_{ij}(r)$ is set to the potential of mean force (PMF) — a crude approximation for true potentials[45] as

$$u^{(0)}_{ij}(r) = -k_B T \ln g^*_{ij}(r) \tag{1}$$

where $g^*_{ij}(r)$ is the experimentally observed pair distribution function and can be calculated from the native structures as follows.

$$g^*_{ij}(r) = \frac{1}{K}\sum_{k=1}^{K} g^{k*}_{ij}(r) \tag{2}$$

where K is the number of the proteins in the training database. $g^{k*}_{ij}(r)$ is the pair distribution function of the k-th native structure and can be calculated as

$$g^{k*}_{ij}(r) = \rho^{k*}_{ij}(r)/\rho^{k*}_{ij,\text{bulk}} \tag{3}$$

where

$$\rho^{k*}_{ij}(r) = \frac{n^{k0}_{ij}(r)}{4\pi r^2 \Delta r} \quad \text{and} \quad \rho^{k*}_{ij,\text{bulk}} = \frac{N^{k0}_{ij}}{V(R_{\max})} \tag{4}$$

are the densities of the ij atom pairs from non-neighboring residues in a spherical shell of radius from $r - \Delta r/2$ to $r + \Delta r/2$ and in a reference sphere of radius $R_{\max}$, respectively. $n_{ij}^{k0}(r)$ is the number of ij pairs in the spherical shell at distance r and N_{ij}^{k0} is the total number of ij pairs in the reference sphere for the k-th native structure. In the present work, the bin size Δr was set to $0.2\,\mathring{A}$ and the radius of the reference sphere $R_{\max}$ was set to $15\,\mathring{A}$.

During the iteration, the pair distribution functions $g_{ij}^{(n)}(r)$ predicted with a set of trial potentials $u_{ij}^{(n)}(r)$ at the n-th iteration cycle can be calculated by using statistical mechanical principles[45] as follows. We consider each protein conformation (l) as a microstate, and consider the native structure ($l = 0$) and the pre-generated decoy conformations ($l = 1,2,3,\cdots,L$) for the k-th protein of the training set as the k-th subsystem. The partition function for the k-th protein (or subsystem) is defined as

$$Z_k = \sum_{l=0}^{L} e^{-\beta E_k^l}, \quad \text{with } E_k^l = \sum_{ij}^{r < r_{\text{cut}}} u_{ij}^{(n)}(r) \tag{5}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, and T is the absolute temperature. E_k^l is the folding energy of the l-th conformation/microstate for the k-th protein, and the summation is over all possible atom pairs from non-neighboring residues in the protein conformation within a distance cutoff r_{cut}. Thus, the probability for the k-th protein to occupy the l-th microstate/conformation is

$$P_k^l = \frac{e^{-\beta E_k^l}}{Z_k} \tag{6}$$

Then, we can calculate the average pair distribution function $g_{ij}^{(n)}(r)$ of atom pair ij for the whole system (i.e., all the proteins in the training set) that consists of K proteins as

$$g_{ij}^{(n)}\left(r\right) = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{kl}\left(r\right) \tag{7}$$

where $g_{ij}^{kl}(r)$ is the pair distribution function for atom pair ij observed in the l-th conformation/microstate of the k-th protein and can be calculated similar to Eq. (3). Apparently, $g_{ij}^{kl}(r)\big|_{l=0} = g_{ij}^{k^*}(r)$.

Next, we use an iterative process to optimize the pairwise potentials $u_{ij}(r)$. Specifically, using the initial trial potentials $u_{ij}^{(0)}(r)$, we calculate the (predicted) average pair distribution functions $g_{ij}^{(0)}(r)$ using Eqs. (5–7). Notice that $g_{ij}^{(0)}(r)$ involves both native and decoy structures.

$g_{ij}^{(0)}(r)$ is a function of $u_{ij}^{(0)}(r)$, which determines the predicted occupancy of each conformer state.

Due to the considerable difference between $u_{ij}^{(0)}(r)$ and (true) $u_{ij}^{*}(r)$, the calculated $g_{ij}^{(0)}(r)$ is expected to significantly differ from (true) $g_{ij}^{*}(r)$. Then, we introduce the following potential correction term to $u_{ij}^{(0)}(r)$:

$$\Delta u_{ij}^{(0)}(r) = \lambda k_{\mathrm{B}} T\left[\ln g_{ij}^{(0)}(r) - \ln g_{ij}^{*}(r)\right] \tag{8}$$

where λ represents a convergence parameter with $0 < \lambda \leq 1$. Here, λ is set to the optimized value (e.g. 0.5 in the present study) for an optimal convergence. With the corrections, we obtain a set of improved potentials as

$$u_{ij}^{(1)}(r) = u_{ij}^{(0)}(r) + \Delta u_{ij}^{(0)}(r) = u_{ij}^{(0)}(r) + \lambda k_{\mathrm{B}} T\left[\ln g_{ij}^{(0)}(r) - \ln g_{ij}^{*}(r)\right] \tag{9}$$

To avoid possible numerical overflow due to the logarithm function log() at low/zero atom-pair frequency at some distances for $g_{ij}(r)$, we replace $\left[\ln g_{ij}^{(0)}(r) - \ln g_{ij}^{*}(r)\right]$ with $\left[g_{ij}^{(0)}(r) - g_{ij}^{*}(r)\right]$ as follows:

$$u_{ij}^{(1)}(r) = u_{ij}^{(0)}(r) + \lambda k_{\mathrm{B}} T\left[g_{ij}^{(0)}(r) - g_{ij}^{*}(r)\right] \tag{10}$$

The second reason for this replacement is to avoid amplifying the sparse data errors by the logarithm function at low pair frequencies.

After each iteration, the pair potentials were truncated by setting $u_{ij}(r) = u_{ij}(r) \times (r_{\mathrm{cut}} - r)/(r_{\mathrm{cut}} - 10)$ for $10\ \text{Å} \leq r \leq r_{\mathrm{cut}}$ so that the potentials gradually approach zero at the cutoff of r_{cut}. Repeating the above iterative procedure (step $n = 1,2,\cdots$), the pair potentials $u_{ij}^{(n)}(r)$ converge to the effective pair potentials $u_{ij}^{*}(r)$.

Using the derived effective pairwise potentials $u_{ij}^{*}(r)$, our new energy scoring function for structure prediction takes the following form:

$$\text{energy score} = \sum_{ij} u_{ij}^{*}\left(r\right) \tag{11}$$

2.2 The statistical-mechanical basis of the iterative method

In the last section, we have presented an iterative method to derive the effective pair potentials $u_{ij}^{*}(r)$. The reason why the resulting potentials $u_{ij}^{*}(r)$ can discriminate the native

structures from the decoys after the pair distribution function converges (i.e. $g_{ij}^{(n)}(r) \rightarrow g_{ij}^{*}(r)$) is explained as follows.

As aforementioned, at the end of the iteration, $g_{ij}^{(n)}(r)$, written as $g_{ij}(r)$ for short, converges to the experimentally observed pair distribution function $g_{ij}^{*}(r)$:

$$g_{ij}(r) \simeq g_{ij}^{*}(r) \tag{12}$$

Substituting the expressions of $g_{ij}(r)$ in Eq. (7) and $g_{ij}^{*}(r)$ in Eq. (2) into the above equation, we have

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{kl}\left(r\right) \simeq \frac{1}{K}\sum_{k=1}^{K} g_{ij}^{k*}\left(r\right) = \frac{1}{K}\sum_{k=1}^{K} g_{ij}^{k0}\left(r\right) \tag{13}$$

where

$$P_k^l = \frac{e^{-\beta E_k^l}}{Z_k}, \quad E_k^l = \sum_{ij}^{r<r_{\text{cut}}} u_{ij}^{*}\left(r\right) \tag{14}$$

Because the sum of the probabilities $\sum_{l=0}^{L} P_k^l = 1.0$, the right-hand side of Eq. (13) can be expressed as

$$\frac{1}{K}\sum_{k=1}^{K} g_{ij}^{k0}\left(r\right) = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{k0}\left(r\right) \tag{15}$$

Substituting Eq. (15) into Eq. (13), we have

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{kl}\left(r\right) \simeq \frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{k0}\left(r\right) \tag{16}$$

Then, the above equation becomes

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l\left[g_{ij}^{kl}(r) - g_{ij}^{k0}(r)\right] = \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L} P_k^l\left[g_{ij}^{kl}(r) - g_{ij}^{k0}(r)\right] \simeq 0 \tag{17}$$

One solution to Eq. (17) is

$$P_k^l\big|_{l=1,2,3,\cdots,L} \simeq 0 \tag{18}$$

According to Henderson's uniqueness theorem for fluid systems,[57] if a set of effective pair potentials $u_{ij}^*(r)$ between atom types i and j can reproduce the pair distribution functions $g_{ij}^*(r)$ of a system, $u_{ij}^*(r)$ is a unique solution. Assuming the uniqueness theorem holds in the complex protein systems, the potentials would be unique if the pair potentials $u_{ij}^*(r)$ reproduce the pair distribution $g_{ij}^*(r)$ of the system. Since P_k^l is determined by $u_{ij}^*(r)$, the solution given by Eq. (18) would be the unique solution to Eq. (17). We therefore have

$$P_k^l\big|_{l=0} = \sum_{l=0}^{L} P_k^l - \sum_{l=1}^{L} P_k^l \simeq 1.0 \tag{19}$$

Substituting Eq. (19) into Eq. (14), we have

$$P_k^0 = \frac{e^{-\beta E_k^0}}{Z_k} \simeq 1.0 \tag{20}$$

for the system at the end of the iteration. Eq. (20) can be rewritten as

$$Z_k \simeq e^{-\beta E_k^0} \tag{21}$$

Combining Eqs. (14) and (21), we have

$$P_k^l\big|_{l=1,2,3,\cdots,L} = \frac{e^{-\beta E_k^l}}{Z_k} \simeq \frac{e^{-\beta E_k^l}}{e^{-\beta E_k^0}} = e^{-\beta\left(E_k^l - E_k^0\right)} \tag{22}$$

Substituting Eq. (22) into Eq. (18), we have

$$e^{-\beta\left(E_k^l - E_k^0\right)} \simeq 0 \tag{23}$$

Re-organizing Eq. (23) gives

$$-\beta\left(E_k^l - E_k^0\right)\big|_{l=1,2,3,\cdots,L} \ll 0 \tag{24}$$

The above equation can be rewritten as

$$E_k^0 \ll E_k^l\big|_{l=1,2,3,\cdots,L}, \quad \text{where } E_k^l = \sum_{ij}^{r<r_{\text{cut}}} u_{ij}^*(r) \tag{25}$$

Namely, if the pair distribution function converges, i.e. $g_{ij}(r) \to g_{ij}^*(r)$, the resulting effective potentials $u_{ij}^*(r)$ will be able to discriminate the native structures from the decoys by the calculated energies. As shown in Eq. (25), $u_{ij}^*(r)$ will yield a much lower energy score to the native structure (E_k^0) than to the decoys $(E_k^l|_{l=1,2,3,\cdots,L})$, which is consistent with the commonly-accepted funnel theory of protein folding.[58–60]

2.3 Protein database used for iteration

The high-resolution decoy set prepared by Floudas and colleagues[36] were used in the present study. The set contains 1225 non-homologous proteins that were selected by Zhang and Skolnick[61]. Each protein has 500~1600 decoy conformations that were generated with the NMR structure refinement program DYANA[62] by retaining distance information among the residues within the hydrophobic core of the protein.[36,63] To exclude the possible effect of the homologous proteins, we removed the proteins that have a sequence similarity above 35% to any protein in the test sets, yielding a total of 1201 non-homologous proteins to be used in our iterative procedure. To save the computer memory, we took every other decoys (i.e. the decoys with odd identification numbers) for the iteration. Therefore, each protein in our training set has one native structure and up to 800 decoys.

Only the heavy atoms are considered in our scoring function. To increase the statistics of the atomic pairs, the 167 heavy atoms in the 20 standard protein residues are grouped into 20 atom types, following their physicochemical properties and chemical environments[64] (Table 1). Our large training database results in significant statistics of frequencies >2000 for all of the 210 possible pairs by the 20 atom types. For example, the frequency is 29,618,116 for C3C-C3C pair. The huge number of pair occurrences for most atom pairs in the training set warrants sufficient statistics to derive the distance-dependent pair potentials despite the limited number of decoys for each protein. Moreover, the global, physics-based iterative function we used which improves the potentials through the comparison of the predicted and observed pair distribution functions poses significant constraints on the degrees of freedom in the parameter space and therefore reduces the actual degrees of freedom, compared to the empirical mathematical functions used by general parameter optimization methods. Finally, our use of potentials of mean force as the initial guess of the pair potentials further helps the fast convergence of iterations for deriving the potentials.

To test the dependence of the derived potentials on the training set, we also used all the 1225 proteins and the other set of decoys that have even identification numbers for our iterative procedure. The results are described in Supporting Information. The derived potentials resulted in little difference for the test sets, which is consistent with our previous finding that the presence of a small portion of homologous proteins/complexes in the training set

compared to the test sets would not significantly change the derived potentials and the test results.[64] The robustness of our iteration method can also be demonstrated by the fact that the derived potentials based on two different sets of decoys yielded no significant results on the test sets (Supporting Information).

2.4 Test sets to evaluate our scoring function

Numerous decoy sets have been constructed for evaluation and development of scoring functions in protein structure prediction. They serve as different benchmarks, according to the algorithms used for decoy generation. The decoy sets in which the structures are fully minimized and the bonds and torsions are fully relaxed are ideal to test scoring functions that consider only non-bonded potentials. Other decoy sets such as RosettaAll,[12] RosettaTsai,[65] and four sets from the Decoys ' *R* ' Us[66] including 4state,[67] lmds,[68] fisa,[12] and vhp_mcdm[69] are designed to test those scoring functions with both the bond-related (e.g. stretching and torsional) and non-bonded interactions; in these test sets some conformations exhibit high stretching/torsional energies.[70] Improper selection of the test sets may cause a bias to the evaluation of a scoring function.[70] Considering that the present scoring function consists of the non-bonded pair potentials, we used the following three decoy sets with sufficiently relaxed torsions:

The first test set is the high resolution (HR) decoy datasets prepared by Floudas and colleagues.[36] It includes 148 non-homologous proteins with 500~1600 high resolution (HR) decoys for each protein. Most of the decoys have an rmsd less than 6~7 Å. This set is to test the ability of a scoring function to distinguish between similar structures with low rmsds, which has useful application to structural refinement to obtain a high-resolution protein structure for drug design. Both this test set of 148 proteins and our training set of 1225 proteins belong to a well represented collection of 1489 non-homologous proteins with a sequence similarity cutoff of 35% prepared by Zhang and Skolnick.[61]

The second test set is the AMBER benchmarking decoy set prepared by Wroblewska and Skolnick,[26] which were generated via AMBER molecular dynamic simulations. The set consists of 47 non-homologous proteins with lengths from 41 to 200 residues below 35% sequence similarity. Each protein includes 1040 decoy structures that are minimized snapshots from AMBER/GBSA molecular dynamics simulations.

The third test set is the CASP8 server decoys for 123 proteins. All the decoys generated by the servers participated in CASP8 were downloaded from the official site of CASP8 (http://predictioncenter.org/). Only those decoys with full length prediction were considered. We also deleted those residues in the decoys that do not present in the native structure for comparability. This yielded a total of 25003 decoys, with an average of 203 decoys per protein.

3 Results

3.1 Extracted potentials

With the iterative method described in the Materials and Methods, we derived a set of 210 effective pair potentials on the basis of 20 protein atom types (Table 1). During the

iteration, the potentials of mean force were used as the initial guess of the potentials as shown in Eq. (1) and the cutoff distance for the potentials was set to $12.\mathring{A}$ To reduce the local correlation effects of covalent bonds, only the interatomic interactions between non-neighboring residues were considered when evaluating the energy score of a protein structure. After each iterative cycle, say the n-th step, the following convergence criterion was checked

$$\Delta g_{\max}^{(n)} = \max\left\{\Delta g_{ij}^{(n)}\right\} \leq \eta \text{ where } \Delta g_{ij}^{(n)} = \frac{1}{S}\sum_{s=1}^{S}\left|g_{ij}^{(n)}(r_s) - g_{ij}^{*}(r_s)\right| \text{ and } i, j = 1,2,\cdots,20 \tag{26}$$

Here, $g_{ij}^{(n)}$ are the predicted pair distribution functions at the n-th iterative cycle. $S = R_{\max}/\Delta r$ is the number of the distance bins for calculating $g_{ij}^{(n)}$, where $R_{\max}$ is the radius of the reference sphere and Δr is the bin size. Details are described in the Materials and Methods and Figure 1. The convergence parameter η was set to 0.01. Our iterative procedure converges rapidly within 20 steps.

Figure 2 shows a selected set of the derived potentials. Several notable features can be observed from the figure, which are physicochemically consistent with experimental findings. For the atom pairs of Car-Car and C3C-C3C, the interaction potentials are favorable around 4 Å, agreeing with the hydrophobic interactions between these atom types. The Car-Car interaction is slightly stronger at a shorter distance than the C3C-C3C pair because the Car-Car pair is involved in an additional aromatic $(\pi - \pi)$ interaction. For the N2C-OC, N3C-OC, N2N-O2M, and O3H-O3H pairs, there exists a minimum between $2.8\mathring{A}$ and 3.0 Å because these atom pairs may form hydrogen bonds. The N2C-OC and N3C-OC pairs shows a stronger and wider interaction than the N2N-O2M and O3H-O3H pairs, because the two atom types in the N2C-OC and N3C-OC pairs are oppositely charged and result in an additional, favorable electrostatic interaction. It is also reasonable that the N3C-OC interaction is stronger than the N3C-OC interaction because the atom type N3C carries more partial charges than N2C.

3.2 Validation of our scoring function

3.2.1 Test set 1: the high resolution decoys generated by Rajgaria et al.

—We first evaluated our scoring function, referred to as ITScore/Pro, using the high resolution (HR) decoy sets constructed by Rajgaria et al.,[63] which includes a total of 148 proteins with 500~1600 high resolution decoys for each protein. The performance is summarized in Table 2, in which six criteria were adopted for evaluation. For reference purpose, the table also lists the results of nine published scoring functions for protein structure prediction,[34,36,48,63,71–74] our previously derived scoring function ITScore/PP for protein-protein interactions,[64] and the potentials of mean force (PMF) derived from an atom-randomized reference state.

It can be seen from Table 2 that our scoring function ITScore/Pro obtains a significant improvement compared to the other scoring functions. Of six assessment parameters,

ITScore/Pro achieves the best performance in five criteria except in Z-Score (the best Z-Score value of 6.02 is achieved by dDFIRE). Specifically, of the 148 proteins, our scoring function identifies 146 native structures as rank #1, yielding a success rate of 98.7% if only the top structure is considered for each protein (Table 2, Column 3).

A second useful scoring criterion is the correlation coefficient (CC) between the energy score and the rmsd of decoys, which is an indicator of the ability of a scoring function to refine a model from a high-rmsd structure to a low-rmsd conformation.[27] On this aspect, ITScore/Pro gives the highest average score-rmsd correlation coefficient (CC=0.82) with the 148 proteins (Table 2, Column 5). Detailed analysis shows that 71.0% of the proteins have a CC>0.8 and 91.2% have a CC>0.6. The score vs. rmsd relationships for 12 example proteins selected by a random number generator program are plotted in Figure 3.

A third notable feature in Table 2 is that our scoring function gives the lowest average rank (1.05) for the native structures (see Column 2). Ideally, a perfect scoring function should achieve an average rank of 1, representing that all the native structures are top ranked compared with their respective decoy structures. Detailed analysis of our results show that there are only two proteins whose native structures have a rank >1 (i.e., 7 for 1g1xC and 2 for 2u1a_).

Finally, it is noticeable in Table 2 that the performance of ITScore/PP ranks the second and is similar to the performance of ITScore/Pro for this test set. It is not surprising that ITScore/Pro and ITScore/PP share some similarities as they both characterize non-covalent interactions among protein atoms. However, ITScore/PP, which was derived from protein-protein complexes, is not optimized for structure prediction of isolated proteins (see below).

3.2.2 Test set 2: the AMBER benchmarking decoy set—Next, we tested our scoring function using the AMBER benchmarking decoy set constructed by Wroblewska and Skolnick.[26] The test set includes 47 proteins, each of which has 1040 decoys. In the original study,[26] this decoy set was designed to test the ability of the AMBER/GBSA force field[75–77] in distinguishing the native structures from decoys. Both the native structures and the decoys were relaxed by 2ns of molecular dynamic simulations with AMBER/GBSA. The decoy set is therefore challenging because all the atoms in each conformation form good atomic contacts through the AMBER simulation.

Figure 4 shows the success rate of our scoring function ITScore/Pro in discriminating the native structures from the decoys. For validation purpose, the results of the scoring functions PMF, AMBER/GBSA, MODELLER/DOPE,[48] DFIRE 2.0,[73] dDFIRE,[74] and ITScore/PP[64] are also shown in the figure. It can been seen that ITScore/Pro yields a success rate of 55.3% in recognizing the native structures as the top rank, vs 8.5% for ITScore/PP,[64] 20% for AMBER/GBSA[26], 29.8% for DFIRE 2.0,[73] 34% for MODELLER/DOPE,[48] 42.6% for PMF, and 57.5% for dDFIRE.[74] This suggests the potential use of our scoring function in structure refinement.

The fact that ITScore/PP performs significantly worse than ITScocre/Pro on the AMBER benchmarking decoy set indicates that the two scoring functions cannot replace each

other. The difference in the two sets of potentials might arise from different degrees of solvation and entropic effects at the protein surface and inside the protein core. In addition, the structure of an isolated protein is fully optimized, whereas the protein-protein interface is not necessarily and depends on the surface structures of each protein partner. Therefore, ITScore/PP performs better for predicting protein-protein complex structures and ITScore/Pro works better for isolated globular proteins.

3.2.3 Test set 3: the CASP8 server decoys—We further applied our scoring function ITScore/Pro to a more realistic test case - the decoy set of 123 proteins downloaded from the server predictions at the CASP8 site (http://www.predictioncenter.org/). Considering that the experimental structures are not known during CASP competitions, the native structures were not included in test set 3. For an automated structure prediction algorithm based on energy ranking, a commonly-used index is the correlation between the scores and the native structural similarity (e.g., rmsd) of the generated decoys.[27] A significant correlation is requisite for a scoring function to rank the decoys and select a protein conformation close to the native structure during the conformational search.

Table 3 lists the results of ITScore/Pro with the CASP8 decoy set. For benchmark purpose, we also showed the results of PMF, MODELLER/DOPE,[48] DFIRE 2.0,[73] dDFIRE,[74] and ITScore/PP[64] in the table. It can be seen that our scoring function ITScore/Pro yields a good score-rmsd correlation with an average Pearson coefficient of 0.672, compared to 0.634 for DFIRE 2.0, 0.595 for DOPE, 0.562 for ITScore/PP, 0.555 for dDFIRE, and 0.394 for PMF. The improvement is more significant when considering the percentage of the proteins in the CASP8 decoy set that receive a high correlation coefficient (i.e., above 0.8): The percentage is 40.7% for ITScore/Pro, vs 26.8% for DFIRE 2.0, 22.0% for DOPE, 17.9% for ITScore/PP, 0.07% for dDFIRE, and 0.02% for PMF, respectively (Figure 5). Correspondingly, the top selected models from our scoring function have a better quality, with an average TM-Score of 0.600 and GDT_TS score of 0.517 (Table 3). TM-Score and GDT_TS are measurements of the similarity between a mode and the native structure.[78]

It can also be seen from Table 3 that the best models in CASP8 were not identified by ITScore/Pro. To test whether the derived potentials score mainly the level of compactness of the structural models, we introduced a control scoring function that uses simple contact-based potentials. In the control, the interaction potential was set to −1.0 (favorable) for an atom pair within a distance of 5 Å, and 0 (no interaction) otherwise. The definition of atom pairs was the same as the definition for ITScore/Pro. The contact-based score is expected to be proportional to the compactness of the structure. Table 3 shows that the contact potentials yielded significantly worse results than the other scoring functions, implying the physics behind the derived potentials of ITScore/Pro.

Again, ITScore/Pro is found to perform better than ITScore/PP for the CASP8 decoy set, indicating the difference between surface potentials and intra-globular potentials.

3.3 How the reference state affects the pairwise potentials

To test how a randomized reference state may affect the predictions, we developed a knowledge-based scoring function referred to as PMF which uses an atom-randomized

reference state. We applied PMF to the three test sets. The results are shown in Table 2, Table 3, Figure 4 and Figure 5. PMF did not perform as well as the other knowledge-based scoring functions (ITScore/Pro, MODELLER/DOPE, DFIRE 2.0, dDFIRE and ITScore/PP), as ITScore/Pro and ITScore/PP circumvent the reference state problem through a statistical mechanics-based iteration method, and DFIRE 2.0, dDFIRE, and MODELLER/DOPE introduce a corrected reference state.[48]

Then, an intriguing question is how and why the choice of the reference state makes a difference on the performances of knowledge-based scoring functions. Elucidation of the underlying mechanism could provide valuable information for the development of other physical scoring functions. To address this question, it is necessary to compare pairs of interaction potentials between our new scoring function and PMF. However, direct comparisons of every pair of interaction potentials are neither practical nor interpretable, because there are 20 atom types and therefore 400 distance-dependent interaction pair potentials. In the present study, we developed the following strategies.

First, as each derived interaction potential is a continuous function of the inter-atom distance, for simplicity, we used a single parameter ϵ_{ij} to roughly represent each potential where ϵ_{ij} is the well depth. ϵ_{ij} reflects the interaction strength between the atom pair ij.

Next, we introduced the following sensitivity variable for each atom type (say, type) in terms of standard deviation to characterize how sensitive the pairwise interaction potentials (represented by ϵ_{ij}) are to different atom type j :

$$\sigma_i = \sqrt{\frac{1}{N_{\text{typ}}} \sum_{j=1}^{N_{\text{typ}}} (\epsilon_{ij} - \bar{\epsilon}_i)^2}, \quad \bar{\epsilon}_i = \frac{1}{N_{\text{typ}}} \sum_{j=1}^{N_{\text{typ}}} \epsilon_{ij} \tag{27}$$

where N_{typ} is the total number of atom types in a scoring function. A knowledge-based scoring function with interaction potentials that exhibit poor atom sensitivity is not expected to be helpful to protein structure prediction.

Figure 6 shows the selectivity parameters σ_i of the twenty atom types in our new scoring function. For comparison, the figure also lists the corresponding selectivity parameters for PMF. Several common characteristics can be observed from the figure, which are consistent with the experimental findings: Both our model and PMF show low selectivity for the nonpolar atom types such as C2M, C2S, Car, C3C, and C3A because these carbon atom types are mainly involved in the non-selective van der Waals interactions. The polar atom types such as N2C, N3C, O2C, and S31 exhibit high selectivity because they are involved in strong electrostatics interactions and/or direction-dependent hydrogen bonding. The atom types C2+ and C2− also manifest higher selectivity than other carbon atom types because C2+ and C2− belong to selective charged groups. However, overall speaking, Figure 6 shows that our model have significantly higher atom selectivity than PMF. In other words, the reference state may alter the performance of a knowledge-based scoring function by

changing the selectivity of its potentials. The finding of the difference in the atom selectivity of the potentials may explain the significantly better performance of our model than PMF.

4 Conclusion and Discussion

In the present study, we have presented a statistical mechanics-based iterative method to extract distance-dependent, all-atom potentials by utilizing the physical pair distribution functions. With the method, we developed a new scoring function for protein structure prediction. The derived scoring function has been validated using three test sets on its ability to discriminate native/near-native conformations and to rank the decoys. Comparisons between our model and a PMF that uses an atom-randomized reference state have revealed that the better performance of our scoring function may attribute to the higher atom selectivity for the derived interaction potentials in our scoring function than the selectivity for the potentials in PMF. The differences in atom sensitivity could provide useful guidance for developing/improving other knowledge-based and physical scoring functions. Underestimated or over-amplified atom sensitivity in potentials may lead to failure in scoring functions on predictions.

Future studies are outlined as follows. The first issue is to search for the best procedure for decoy generation if one plans to use a larger training database of known protein structures to re-derive the effective potentials. This is also a common problem in deriving potentials based on decoys.[27] The-oretically, two ideal approaches can be used to construct ensembles of decoy conformations for our iterative method, which would give an accurate definition of the system partition function. The first approach is to generate an ensemble of conformations for each protein with the current potentials at every iterative step, using Monte Carlo or molecular dynamics simulations.[56] Then, the pair distribution functions are calculated from the ensembles of conformations generated on the fly. The other approach is to exhaustively generated all the possible folding conformations before iteration.[44] However, both methods are computational impractical because the system contains tens of atom types, thousands of atoms, and thousands of proteins. Performing Monte Carlo simulations at every iterative cycle or exhaustively generating all the possible conformations of a protein are both beyond the current computational power. A future procedure for decoy generation should achieve the following three requirements: First, the generated decoy conformations should be well sampled so as to cover the entire protein conformational space. Second, the correlation between native similarity and radius of gyration should be small.[27] Otherwise, the native structures could be easily identified in the first iterative step and no adjustment will be made for the potentials. Third, as our iterative method does not include bonding and torsional potentials, the generated decoys should contain no distorted bonds or torsional angles. Otherwise, the conformational artifact effect may be introduced into the pair potentials due to the high bonding or torsional energies.

The second issue for improvement is to include the bond-related potentials (such as bonding or torsional energies) in the present non-bonded pair potentials. As shown in the Results Section, the success rate of our scoring function for the HR decoy sets is significantly higher than that for the AMBER benchmarking decoys. Part of the reasons may be that the AMBER benchmarking decoys are generated via AMBER simulations[26] in which some

of the bonds could be distorted to form good atomic contacts, leading to higher bonding or torsional energies. However, no bonding or torsional energy penalties are considered in our present scoring function, leading to artificially lower energy scores for some distorted decoys. Moreover, the bond-related energies are also important in structure refinement.[27] Therefore, it will be necessary to include the bonding and torsional potentials in future studies.

Thirdly, as shown in the derivations using statistical mechanical principles (see Materials and Methods), our iterative method warrants that the energy scores of native structures are considerably lower than those of the decoys by using the convergence of the pair distribution functions. A future study could be to introduce an additional optimization process in our iterative procedure to maximize the correlation of the energy score with the native similarity such as rmsd or TM-score,[27,28] though this new process may make the potential derivation more artificial and less physical.

Other future directions to improve the present scoring function include considering multi-body (e.g. three-body and four-body) potentials,[80–83] orientational dependence of inter-atomic interactions,[18,49,74] and conformational entropy effects.[84]

It should also be noted that the present pair potentials are effective interaction potentials between atom pairs, rather than the free energies. They represent the overall effect of different energy components such as VDW energies, electrostatic energies, hydrogen bonding, solvation and entropy. However, some energy terms such as solvation and entropy are not pairwise additive and thus the pair potentials may not effectively incorporate the effects of these energies. Therefore, accounting for solvation and entropy in addition to the pair interaction potentials may improve the scoring performance.[79]

Finally, although ITScore/Pro is proposed for the purpose of structural selection in protein structure prediction, it would be interesting to test how good ITScore/Pro performs in protein folding simulations. The latter is a more challenging task, which requires correct ranking of not only the lowest energy state but also the full energy landscape.

It is emphasized that all the tests performed in the present study are used for validation of our scoring functions. The results do not serve for competing interests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support to XZ from OpenEye Scientific Software Inc. (Santa Fe, NM) is gratefully acknowledged. XZ is supported by NIH grant R21GM088517, NSF CAREER Award DBI-0953839, the Research Board Award RB-07-32 and the Research Council Grant URC 09-004 of the University of Missouri. The computations were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000; 28:235–242. [PubMed: 10592235]
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2008;36:D25–D30. [PubMed: 18073190]
3. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287. [PubMed: 10700142]
4. Baker D Protein structure prediction and structural genomics. Science 2001;294:93–96. [PubMed: 11588250]
5. Jacobson M, Sali A. Comparative protein structure modeling and its applications to drug discovery. in Annual Reports in Medicinal Chemistry, ed Overington J (Inpharmatica Ltd., London), 2004;39:259–276.
6. Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. Nucleic Acids Res 2005;33:1874–1891. [PubMed: 15805122]
7. Zhou H, Skolnick J. Protein structure prediction by pro-Sp3-TASSER. Biophys J 2009;96:2119–2127. [PubMed: 19289038]
8. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–738. [PubMed: 20360767]
9. Petrey D and Honig B Protein structure prediction: inroads to biology. Mol Cell 2005;20:811–819. [PubMed: 16364908]
10. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18:342–348. [PubMed: 18436442]
11. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325. [PubMed: 10940251]
12. 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. [PubMed: 9149153]
13. Zhang Y, Arakaki AK, Skolnick J. TASSER: An automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61(S7):91–98. [PubMed: 16187349]
14. 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. [PubMed: 10864507]
15. Shmygelska A, Levitt M. Generalized ensemble methods for de novo structure prediction. Proc Natl Acad Sci USA 2009;106:1415–1420 [PubMed: 19171891]
16. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins 2008;71:1175–1182. [PubMed: 18004754]
17. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145. [PubMed: 10753811]
18. Buchete NV, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 2004;14:225–232. [PubMed: 15093838]
19. Skolnick J. In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol 2006;16:166–171. [PubMed: 16524716]
20. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamic calculations. J Comput Chem 1983;4:187–217.
21. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1998;288:477–487.
22. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
23. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general Amber force field. J Comput Chem 2004;25:1157–1174. [PubMed: 15116359]

24. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688. [PubMed: 16200636]
25. Liwo A, Arłukowicz P, Czaplewski C, Ołdziej S, Pillardy J, Scheraga HA. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. Proc Natl Acad Sci USA 2002;99:1937–1942. [PubMed: 11854494]
26. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. J Comput Chem 2007;28:2059–2066. [PubMed: 17407093]
27. Jagielska A, Wroblewska L, Skolnick J. Protein model refinement using an optimized physics-based allatom force field. Proc Natl Acad Sci USA 2008;105:8268–8273. [PubMed: 18550813]
28. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950. [PubMed: 1004017]
29. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 1985;18:534–552.
30. 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. [PubMed: 2359125]
31. Li X, Liang J. Knowledge-based energy functions for computational studies of proteins. in Computational Methods for Protein Structure Prediction and Modeling, eds Xu Y, Xu D, Liang J (Springer), 2006;1:71–124.
32. Zhu J, Fan H, Periole X, Honig B, Mark AE. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins 2008;72:1171–1188. [PubMed: 18338384]
33. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888. [PubMed: 1404392]
34. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: Results from linear optimization. Proteins 2000;41:40–46. [PubMed: 10944392]
35. Qiu J, Elber R. Atomically detailed potentials to recognize native and approximate protein structures. Proteins 2005;61:44–55. [PubMed: 16080157]
36. Rajgaria R, McAllister SR, Floudas CA. Distance dependent centroid to centroid force fields using high resolution decoys. Proteins 2008;70:950–970. [PubMed: 17847088]
37. Hao MH, Scheraga HA. How optimization of potential function affects protein folding. Proc Natl Acad Sci USA 1996;93:4984–4989 [PubMed: 8643516]
38. Bastolla U, Vendruscolo M, Knapp EW. A statistical mechanical method to optimize energy functions for protein folding. Proc Natl Acad Sci USA 2000;97:3977–3981. [PubMed: 10760269]
39. Mimy LA, Shakhnovich EI. How to derive a Protein Folding Potential? A New Approach to an Old Problem. J Mol Biol 1996;264:1164–1179. [PubMed: 9000638]
40. Huber T, Torda AE. Protein fold recognition without Boltzmann statistics or explicit physical basis. Protein Sci 1998;7:142–149 [PubMed: 9514269]
41. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc Natl Acad Sci USA 1998;95:2932–2937. [PubMed: 9501193]
42. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: How accurate are they? J Mol Biol 1996; 257:457–469. [PubMed: 8609636]
43. Koppensteiner WA, Sippl MJ. Knowledge-based potentials - Back to the roots. Biochemistry (Moscow) 1998;63:247–252. [PubMed: 9526121]
44. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628–11633. [PubMed: 8876187]
45. McQuarrie DA. in Statistical Mechanics, 2000, (University Science Books).

46. Poole AM, Ranganathan R. Knowledgebased potentials in protein design. Curr Opin Struct Biol 2006;16:508–513. [PubMed: 16843652]
47. Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys 2006;46:165–174. [PubMed: 17012757]
48. Shen M-Y, Sali A Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524 [PubMed: 17075131]
49. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 2008;376:288–301. [PubMed: 18177896]
50. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008;71:261–277. [PubMed: 17932912]
51. Huang S-Y, Zou X Mean-force scoring functions for protein-ligand binding. Annu Rep Comput Chem 2010;6:281–296.
52. 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. [PubMed: 12381853]
53. Perez C, Ortiz AR. Evaluation of docking functions for protein-ligand interactions. J Med Chem 2001;44:3768–3785. [PubMed: 11689064]
54. Huang S-Y, Zou X An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J Comput Chem 2006;27:1865–1875.
55. Huang S-Y, Zou X An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem 2006;27:1876–1882. [PubMed: 16983671]
56. Almarza NG, Lomba E. Determination of the interaction potential from the pair distribution function: An inverse Monte Carlo technique. Phys Rev E 2003;68:011202(1–6).
57. Henderson RL. A uniqueness theorem for fluid pair correlation functions. Phys Lett A 1974;49:197–198.
58. Chan HS, Dill KA. Transition states and folding dynamics of proteins and heteropolymers. J Chem Phys 1994;100:9238–9257.
59. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Towards an outline of the topography of a realistic folding funnels. Proc Natl Acad Sci USA 1995;92:3626–3630. [PubMed: 7724609]
60. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195. [PubMed: 7784423]
61. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Nat Acad Sci USA 2004;101:7594–7599. [PubMed: 15126668]
62. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298. [PubMed: 9367762]
63. Rajgaria R, McAllister SR, Floudas CA. Development of a novel high resolution Ca-Ca distance dependent force field using a high quality decoy set. Proteins 2006;65:726–741. [PubMed: 16981202]
64. Huang S-Y, Zou X An iterative knowledge-based scoring function for protein-protein recognition. Proteins 2008;72:557–579. [PubMed: 18247354]
65. Tsai J, Bonneau R, Rohl C, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76–87. [PubMed: 12945051]
66. Samudrala R, Levitt M. Decoys ‘R’ Us: a database of incorrect protein conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401. [PubMed: 10933507]
67. 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. [PubMed: 8627632]
68. 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. [PubMed: 12742025]
69. Fogolari F, Tosatto SCE, Colombo G. A decoy set for the thermostable subdomain from chicken villin headpiece, comparison of different free energy estimators. BMC Bioinf 2005;6:301.

70. Handl J, Knowles J, Lovell SC. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. Bioinformatics 2009; 25:1271–1279. [PubMed: 19297350]
71. Loose C, Klepeis JL, Floudas CA. A new pairwise folding potential based on improved decoy generation and side-chain packing. Proteins 2004;54:303–314. [PubMed: 14696192]
72. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994; 243:668–682. [PubMed: 7966290]
73. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008;17:1212–1219. [PubMed: 18469178]
74. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72:793–803. [PubMed: 18260109]
75. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
76. Tsui V, Case DA. Theory and applications of the Generalized Born solvation model in macromolecular simulations. Biopolymers 2001;56:275–291.
77. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
78. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710. [PubMed: 15476259]
79. Huang S-Y, Zou X Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. J Chem Inf Model 2010;50:262–273. [PubMed: 20088605]
80. Tropsha A, Singh RK, Vaisman II. Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–222. [PubMed: 8811483]
81. Krishnamoorthy B, Tropsha A. Development of a four-body statistical pseudo-potential to discriminate native from nonnative protein conformations. Bioinformatics 2003;19:1540–1548. [PubMed: 12912835]
82. Li X, Liang J. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins 2005;60:46–65. [PubMed: 15849756]
83. Feng YP, Kloczkowski A, Jernigan RL. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins 2007;68:57–66. [PubMed: 17393455]
84. Brady PG, Sharp KA. Entropy in protein folding and in protein-protein interactions. Curr Opin Struct Biol 1997;7:215–221. [PubMed: 9094326]

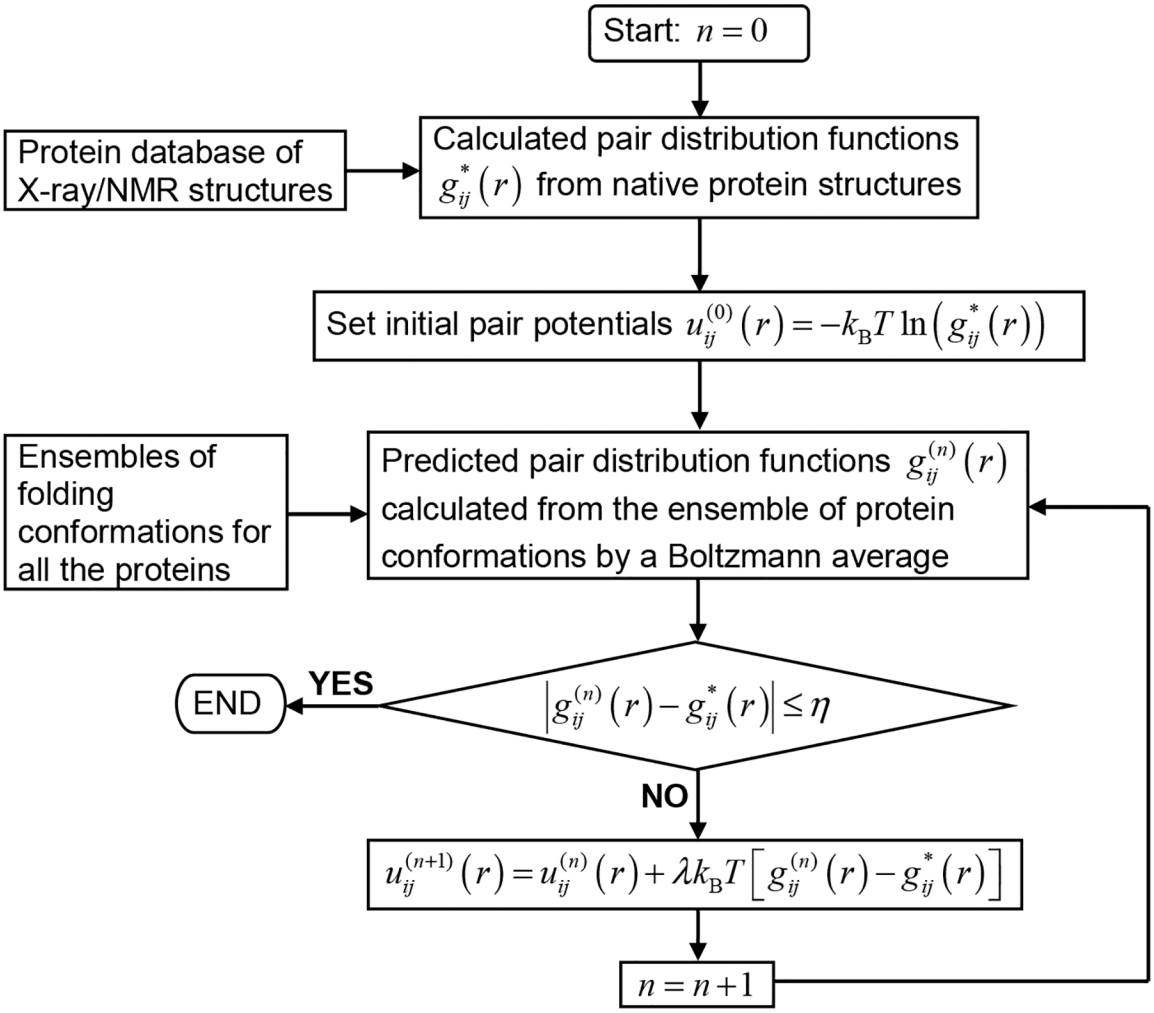

Figure 1:
An illustration of our iterative procedure.

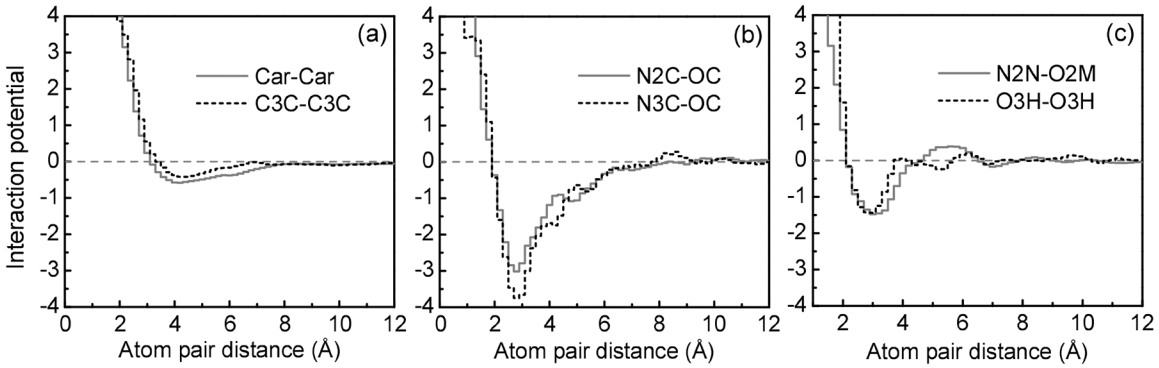

Figure 2:
A set of selected pair potentials extracted by our iterative method.

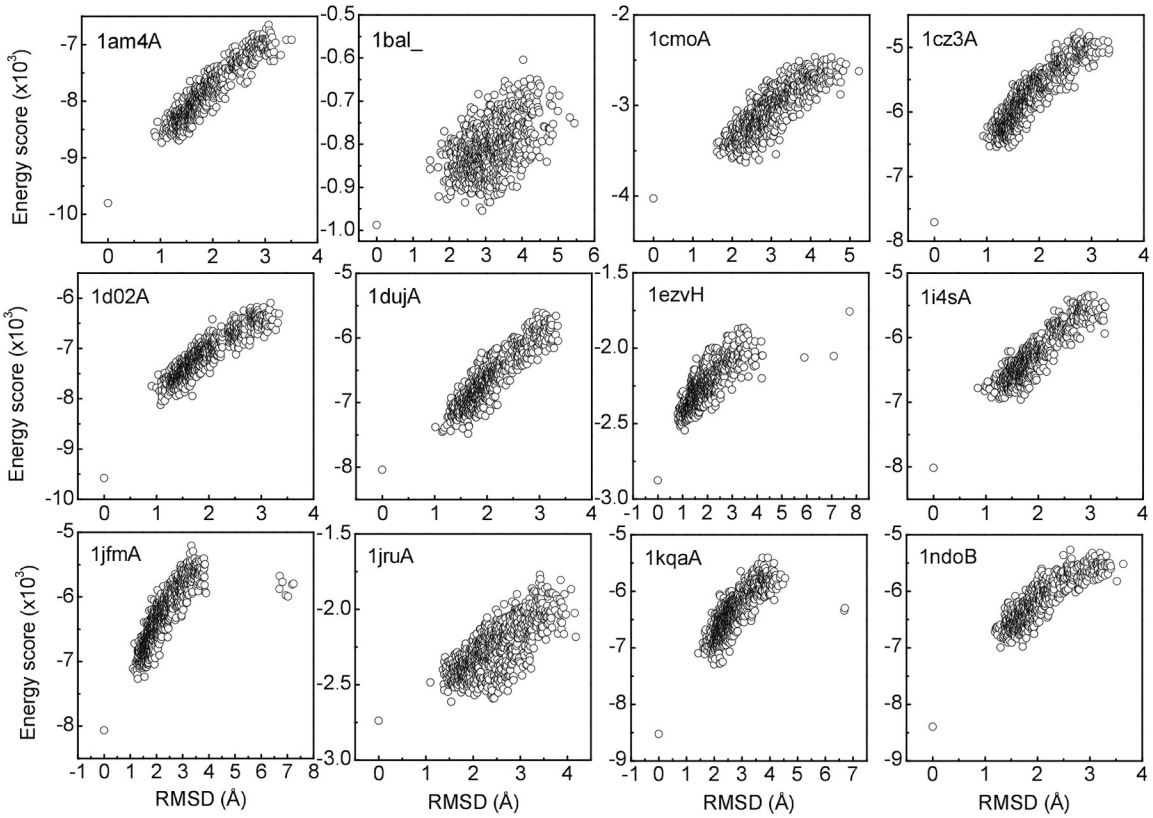

Figure 3:

The score vs. rmsd scatter plots for the decoys of twelve randomly selected proteins in the high resolution (HR) decoy set.

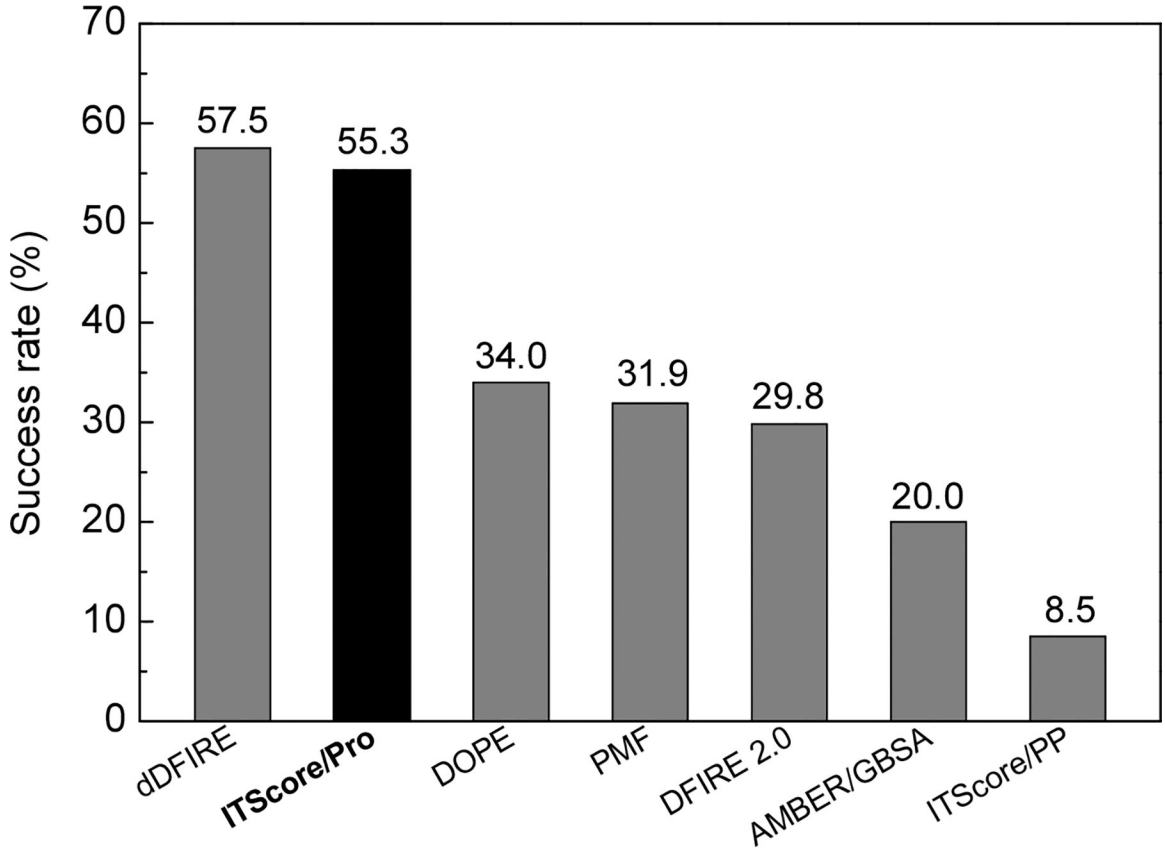

Figure 4:

Success rates of our model ITScore/Pro, DFIRE 2.0,[73] dDFIRE,[74] MODELLER/DOPE,[48] AMBER/GBSA,[75–77], ITScore/PP[64] and PMF on recognizing the native structures for the AMBER benchmarking decoy set prepared by the Skolnick lab. The result for AMBER/GBSA was obtained from the original paper.[26]

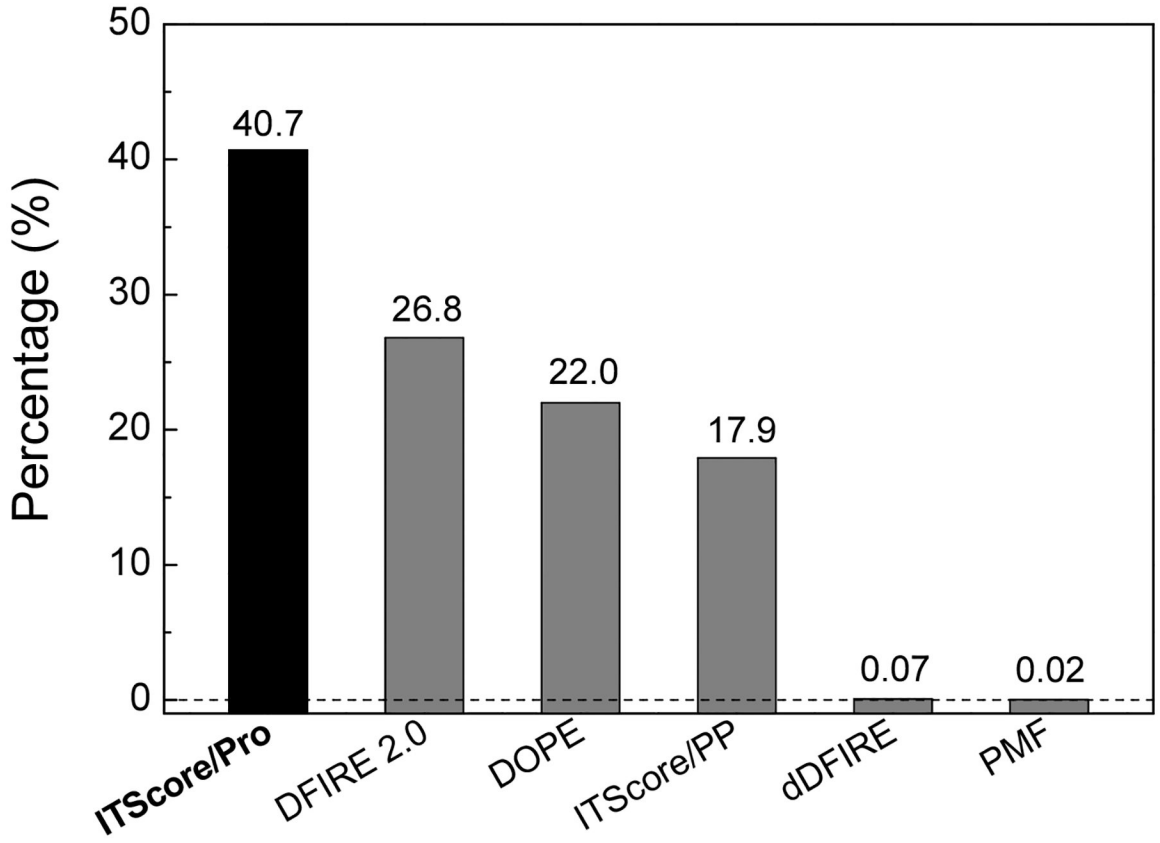

Figure 5:

Percentage of the proteins in the CASP8 decoy set that have a score-rmsd correlation coefficient above 0.8, calculated with our scoring function ITScore/Pro, DFIRE 2.0, dDFIRE, MODELLER/DOPE, ITScore/PP[64] and PMF. See the text for detailed explanations.

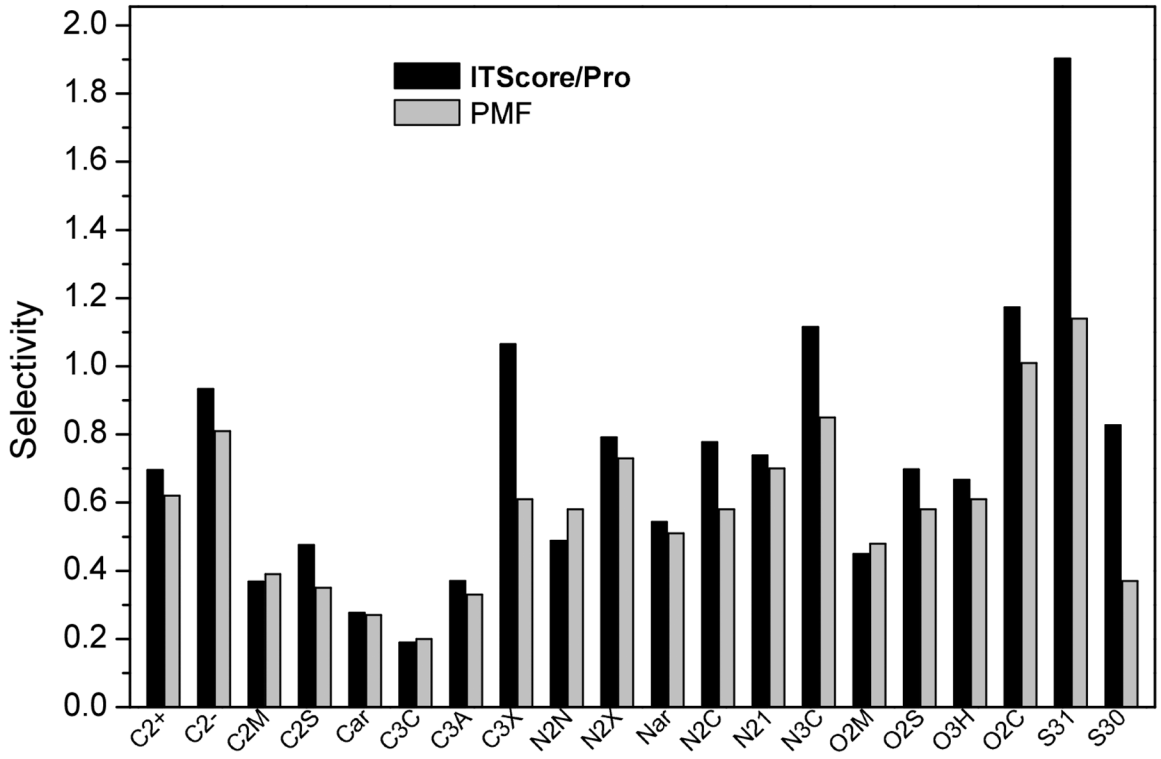

Figure 6:

Comparison of the selectivity parameters σ of the potentials for 20 atom types between our model and PMF.

Table 1:

Twenty atom types for the heavy atoms of proteins.

No.	Symbol	Atom name
1	C2+	ARG_CZ
2	C2−	*_C (on terminal residues), ASP_CG, GLU_CD
3	C2M	*_C (not on terminal residues)
4	C2S	ASN_CG, GLN_CD
5	Car	HIS_CD2, HIS_CE1, HIS_CG, PHE_CD1, PHE_CD2, PHE_CE1, PHE_CE2, PHE_CG, PHE_CZ, TRP_CD1, TRP_CD2, TRP_CE2, TRP_CE3, TRP_CG, TRP_CH2, TRP_CZ2, TRP_CZ3, TYR_CD1, TYR_CD2, TYR_CE1, TYR_CE2, TYR_CG, TYR_CZ
6	C3C	ALA_CB, ARG_CB, ARG_CG, ASN_CB, ASP_CB, GLNL_CB, GLNL_CG, GLU_CB, GLLL_CG, HIS_CB, ILE_CB, ILE_CD1, ILE_CG1, ILE_CG2, LEU_CB, LEU_CD1, LELL_CD2, LEU_CG, LYS_CB, LYS_CD, LYS_CG, MET_CB, PHE_CB, PRO_CB, PRO_CG, SER_CB, THR_CG2, TRP_CB, TYR_CB, VAL_CB, VAL_CG1, VAL_CG2
7	C3A	*_CA
8	C3X	ARG_CD, CYS_CB, LYS_CE, MET_CE, MET_CG, PRO_CD, THR_CB
9	N2N	*_N (not on terminal residues)
10	N2X	*_N (on terminal residues), ASN_ND2, GLN_NE2
11	Nar	HIS_ND1, HIS_NE2, TRP_NE1
12	N2C	ARG_NH1, ARG_NH2
13	N21	ARG_NE
14	N3C	LYS_NZ
15	O2M	*_O (not on terminal residues)
16	O2S	ASN_OD1, GLN_OE1
17	O3H	SER_OG, THR_OG1, TYR_OH
18	OC	*_O or *_OXT (on terminal residues), ASP_OD1, ASP_OD2, GLU_OE1, GLU_OE2
19	S31	CYS_SG
20	S30	MET_SD

"*" stands for any residue.

Table 2:

Summary of the testing results of our scoring function ITScore/Pro and eleven other scoring functions on the high resolution decoy set prepared by Rajgaria et al.[36] The results for HRSC, HR, TE13, HL, and LKF were taken from the original paper[36] and the references therein. The results for DFIRE 2.0[73] and dDFIRE[74] were obtained from our calculations with the executables provided by the Zhou Lab. The results for DOPE[48] were calculated with Modeller 9v1.[11].

Scoring function	Avg rank[a]	No of firsts[b]	Avg Z-score[c]	Avg CC[d]	Avg Cα-rmsd[e]	Avg Cα-rmsd[f]
ITScore/Pro	**1.05**	**146/148 (98.7%)**	**4.48**	**0.82**	**0.028**	**1.64**
ITScore/PP64	1.09	146/148 (98.7%)	4.61	0.76	0.041	2.02
DFIRE 2.073	8.59	142/148 (96.0%)	4.29	0.81	0.095	1.65
dDFIRE74	9.26	140/148 (94.6%)	6.02	0.72	0.117	1.64
DOPE48	18.18	134/148 (90.5%)	4.76	0.72	0.201	1.68
6bin-HRSC36	2.49	128/148 (86.5%)	3.62	0.70	0.298	1.82
7bin-HRSC36	2.01	125/148 (84.5%)	3.39	0.70	0.321	1.83
PMF[g]	48.47	112/148 (75.7%)	3.30	0.40	0.60	1.86
HR63	1.87	113/150 (75.3%)	2.11	0.80	0.451	1.76
TE1334	19.94	92/148 (62.2%)	3.15	0.63	0.813	1.89
HL72	44.93	70/150 (46.7%)	2.34	0.59	1.092	1.84
LKF[71]	39.45	17/150 (11.3%)	1.55	0.52	1.721	1.93

[a] The average rank of the native conformations. The best rank is 1.

[b] The number of the proteins with native structures as rank #1 in terms of the calculated energy scores.

[c] The average Z-score for all the tested proteins, measuring the relative energetic separation of the native structure of a protein with respect to its decoys.

[d] The average Pearson correlation coefficients (CC) between the energy scores and the rmsd of decoys.

[e] The average rmsd of the best predicted structures with the lowest energy scores (native structures included).

[f] The average rmsd of the best predicted structures with the lowest energy scores with native structures excluded.

[g] PMF is a knowledge-based scoring function we derived with an atom-randomized reference state for test purpose. See Section 3.3 for detail.

Table 3:

The performance of our scoring function ITScore/Pro on the test set of the server predictions in CASP8. The results of DFIRE 2.0,[73] dDFIRE,[74] MODELLER/DOPE,[48], ITScore/PP[64], PMF and simple contact-based potentials are also listed as references.

Method	Avg CC[a]	Avg TM-score	Avg GTD_TS
Best model[b]	-	0.677	0.585
ITScore/Pro	**0.672**	**0.600**	**0.517**
DFIRE 2.0	0.634	0.598	0.515
DOPE	0.595	0.583	0.502
ITScore/PP	0.562	0.587	0.504
dDFIRE	0.555	0.586	0.504
PMF	0.394	0.542	0.467
Contact Potentials	0.392	0.419	0.344

[a]The average Pearson correlation coefficients (CC) between the energy scores and the rmsd of the decoys.

[b]For each protein in the CASP8 test set, the best model refers to the decoy structure that is closest to the native structure and thus has the highest TM-score and GTD_TS value compared to other decoy structures of the protein. The average TM-score and GTD_TS values of the best models (first row) represent the highest values that a scoring function can achieve.