PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

What is the best reference state for designing statistical atomic potentials in protein structure prediction?

Haiyou Deng,[1] Ya Jia,[1] Yanyu Wei,[1] and Yang Zhang[2]*

[1] Department of Physics and Institute of Biophysics, Central China Normal University, Wuhan 430079, China

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 45108

ABSTRACT

Many statistical potentials were developed in last two decades for protein folding and protein structure recognition. The major difference of these potentials is on the selection of reference states to offset sampling bias. However, since these potentials used different databases and parameter cutoffs, it is difficult to judge what the best reference states are by examining the original programs. In this study, we aim to address this issue and evaluate the reference states by a unified database and programming environment. We constructed distance-specific atomic potentials using six widely-used reference states based on 1022 high-resolution protein structures, which are applied to rank modeling in six sets of structure decoys. The reference state on random-walk chain outperforms others in three decoy sets while those using ideal-gas, quasi-chemical approximation and averaging sample stand out in one set separately. Nevertheless, the performance of the potentials relies on the origin of decoy generations and no reference state can clearly outperform others in all decoy sets. Further analysis reveals that the statistical potentials have a contradiction between the universality and pertinence, and optimal reference states should be extracted based on specific application environments and decoy spaces.

Proteins 2012; 80:2311–2322.

Key words: protein structure prediction; knowledge-based force field; protein structural decoys; quasi-chemical approximation; native state.

INTRODUCTION

The design and construction of energy function that has a global minimum in the native state are essential for protein folding and protein structure prediction.[1–3] Since Anfinsen's hypothesis [4] was put forward in the 1970s, different types of knowledge-based empirical potentials have developed like mushrooms,[5–7] by virtue of the rapid increase of structure data in the PDB library.[8] Any aspects of structural features that differ substantially between the set of native and nonnative conformations can be used to construct statistical potential,[9] for example, the strength of electrostatic interactions, the torsion angle, the exposure of nonpolar groups to solvent, and so forth. In particular, following the idea of Sippl,[6,10] a variety of atomic-level distance-dependent contact potentials have been recently developed,[9,11–17] and successfully applied to many molecular modeling areas, including fold recognition,[18–20] *ab initio* folding,[21–26] protein structure refinement,[27,28] 3D model assessment,[12,17,29] protein stability analysis,[15,30] and protein–protein docking.[11,31]

Most of the knowledge-based potentials were derived based on the Boltzmann or Bayesian formulations. For the atomic distance-specific contact potentials, the potential can be written as:

$$\bar{u}_{i,j}(r) = -RT \ln \left[\frac{f_{i,j}^{\mathrm{OBS}}(r)}{f_{i,j}^{\mathrm{ERF}}(r)} \right] \qquad (1)$$

where R and T are Boltzmann constant and Kelvin temperature, respectively. $f_{i,j}^{\mathrm{OBS}}(r)$ is the observed probability of atomic pairs (i, j) within a distance bin r to $r+\Delta r$ in

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Natural Science Foundation of China; Grant number: 11175068; Grant sponsor: NSF Career Award; Grant number: DBI 0746198; Grant sponsor: National Institute of General Medical Sciences; Grant number: GM083107, GM084222.

*Correspondence to: Yang Zhang, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 45108.
E-mail: zhng@umich.edu

Received 3 April 2012; Revised 30 April 2012; Accepted 21 May 2012
Published online 24 May 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24121

experimental protein conformations. $f_{i,j}^{\mathrm{REF}}(r)$ is the expected probability of atomic pairs (i, j) in the corresponding distance from random conformations without atomic interactions, which is so-called reference state. Since most existing statistical potentials use the same size of datasets to calculate $f_{i,j}^{\mathrm{OBS}}(r)$ and $f_{i,j}^{\mathrm{REF}}(r)$, the probabilities in Eq. (1) can be replaced by the frequency counts of atomic pairs:

$$\bar{u}_{i,j}(r) \approx -RT \ln \left[\frac{N_{i,j}^{\mathrm{OBS}}(r)/N_{i,j}^{\mathrm{OBS}}}{N_{i,j}^{\mathrm{ERF}}(r)/N_{i,j}^{\mathrm{ERF}}} \right] = -RT \ln \left[\frac{N_{i,j}^{\mathrm{OBS}}(r)}{N_{i,j}^{\mathrm{ERF}}(r)} \right] \tag{2}$$

Here, $N_{i,j}^{\mathrm{OBS}}(r)$ is the observed number of atom pairs (i, j) at the distance r in experimental protein structures. $N_{i,j}^{\mathrm{REF}}(r)$ is the expected number of atomic pairs (i, j) if there were no interactions between atoms. $N_{i,j}^{\mathrm{OBS}} \equiv N_{i,j}^{\mathrm{REF}} = \sum_r^{r_{\mathrm{cut}}} N_{i,j}^{\mathrm{OBS}}(r)$ is the total number of atomic pairs (i, j) in the structure samples, where r_{cut} is the cutoff distance.

The statistical potential in Eqs. (1) and (2) is also known as the potential of mean force. In specific derivations, it needs a clear delineation of distance interval and bin splitting scheme. Meanwhile, it should be clearly defined on what kinds of atoms to be considered, and which set of experimental structures to be used. The most critical step for statistical potentials is the selection of reference states.[2] In principle, the reference state should be obtained from the statistics of random conformations which lacks of inherent atomic interactions and has the ability to offset the statistical biases from specific sample selections and parameter cutoffs.

There is however no universal way as for the construction of the reference states. Common disposal methods for the reference state calculation can be divided into two categories: one is by analytical assumptions, the other is by statistics but the statistical samples are from native protein conformations or their decoys. Because of the importance, a number of studies have been conducted for assessing the performance of different reference states.[2,14,15,17,32] However, because these studies exploited the potentials from the original programs which had been constructed using different databases and programming environments, it remains unclear whether the observed differences in performance is due to the selection of reference state, or due to the technical details of training databases, programming, and parameter cutoffs.

Meanwhile, most of the previous assessment studies were focused on the selection of native structures. Since the native structures can never been generated by computer simulations, a more realistic and challenging task is to prioritize the best *near-native* computer models from the structural decoys. Another critical criterion of the potential development is to examine the correlations of the potential with the similarity to the native (e.g. RMSD, TM-score and GDT_TS),[33] because a better long-range correlation is essential to guide the protein folding simulations from nonnative states to the native ones.[28]

In this article, we made a systematical examination of six most-often used reference states, including averaging,[9] quasi-chemical approximation,[12] finite ideal-gas,[15] spherical noninteracting,[17] atom-shuffled,[16] and random-walk chain.[14] To rule out the dependence of training databases and technical details from original potentials, we reconstructed all the potentials using a uniform dataset by the same programming environment. To establish the generality of the analyses, we applied the potentials to six independent decoy sets, from various resources of template reassembly and *ab initio* folding, with a comprehensive assessment of both native, near-native structure prioritization and energy-TM-score correlation.

METHODS

We constructed six statistical potentials using Eqs. (1) and (2). As in most of previous potential developments, 167 residue-specific heavy atom types are used.[9] The distance cutoff is set to 15 Å with a bin width 0.5 Å, which results in 30 bins. Atom pairs from the same residue are ignored in our pair-wise potential counting. The constructed potential can be written as a 30 × 27,889 matrix. In the cases where certain atom pairs are not observed at specific distance bin, the potentials are set to a score corresponding to the least favorable one in the whole potential.

A unified, nonredundant set of experimental protein structures was collected for the construction of various potentials in this study. The protein list is generated from the PISCES server,[34] with a resolution cutoff 1.6 Å, *R*-factor cutoff 0.25 Å, and sequence identity cutoff 20%. Only the structures determined by X-ray crystallography were considered. In addition, protein structures with incomplete, missing or nonstandard residues were excluded, except for the structures that missed residues only in the terminals. The final sample of the experimental structures contains 1022 protein chains, including 165 α, 100 β, and 713 αβ proteins (others 44 have little secondary structure), which are publicly available at http://zhanglab.ccmb.med.umich.edu/potential/assessment.

The total energy score of a given protein sequence S_{q} with conformation C_{p} is calculated by

$$\mathrm{Score}(S_{\mathrm{q}}, C_{\mathrm{p}}) = \sum_{m=1} \sum_{n=m+1} \bar{u}_{i_m,i_n}(r_{m,n}) \tag{3}$$

where $r_{m,n}$ is the distance between mth and nth atoms, and i_m and i_n are the residue-specific atom types, respectively. m and n runs through all the atoms in the protein chain except for those pairs from the same residues.

Averaging reference state (RAPDF-REF)

The RAPDF potential was proposed by Samudrala and Moult,[9] which uses an average over different atom types in the experimental conformations to represent the random reference states. Therefore, $N_{i,j}^{\mathrm{REF}}(r)$ can be calculated as follows:

$$N_{i,j}^{\mathrm{REF}}(r) = f^{\mathrm{REF}}(r)N_{i,j}^{\mathrm{REF}} = \frac{\sum_{i,j} N_{i,j}^{\mathrm{OBS}}(r)}{\sum_{i,j}\sum_{r} N_{i,j}^{\mathrm{OBS}}(r)} \sum_{r} N_{i,j}^{\mathrm{OBS}}(r) = \frac{N^{\mathrm{OBS}}(r)}{N_{\mathrm{total}}^{\mathrm{OBS}}} N_{i,j}^{\mathrm{OBS}} \quad (4)$$

Here $N^{\mathrm{OBS}}(r)$ is the number of observed contacts between all pairs of atom types at a particular distance r. $N_{\mathrm{total}}^{\mathrm{OBS}}$ is the total number of contacts between all pairs of atom types summed over all distance r. The contact numbers from different proteins in the dataset are pooled together to calculate $N^{\mathrm{OBS}}(r)$, $N_{\mathrm{total}}^{\mathrm{OBS}}$, and $N_{i,j}^{\mathrm{OBS}}$. Here, an assumption of $N^{\mathrm{REF}}(r) = N^{\mathrm{OBS}}(r)$ has been taken by the authors. Thus, $N_{i,j}^{\mathrm{REF}}(r)$ and $N_{i,j}^{\mathrm{OBS}}(r)$ can be derived from the same protein dataset. Although the averaging reference state is easy to calculate, a weakness of the potential is that the contact density distribution for all pairs of atom types is assumed to be the same, which deviates from the reality.

Quasi-chemical approximation reference state (KBP-REF)

In the quasi-chemical approach of Lu and Skolnick,[12] $N_{i,j}^{\mathrm{REF}}(r)$ was defined as:

$$N_{i,j}^{\mathrm{REF}}(r) = x_i x_j N^{\mathrm{OBS}}(r) \quad (5)$$

where x_k is the mole fraction of atom type k, which is calculated based on the whole dataset. Here it has also the assumption $N^{\mathrm{REF}}(r) = N^{\mathrm{OBS}}(r)$. As a reasonable approximation for reference state, the referential number of atomic pairs (i, j) within certain distance bin is proportional to the mole fraction of atom type i and atom type j. The atomic potential using Eq. (5) was named KBP.[12]

Finite ideal-gas reference state (Dfire-REF)

In Dfire potential,[15] Zhou and Zhou exploited a ideal-gas system to simulate the reference state. The number of atom pairs in the system was calculated by:

$$N_{i,j}^{\mathrm{REF,p}}(r) = N_{i,j}^{\mathrm{REF,p}} \frac{4\pi r^2 \Delta r}{V^p} = n_i^{\mathrm{p}} n_j^{\mathrm{p}} \frac{4\pi r^2 \Delta r}{V^{\mathrm{p}}} \quad (6)$$

where V^P is the volume of protein P, n_i^{p} and n_j^{p} are the number of atoms of type i and j in the protein, respectively. Since Eq. (6) is from liquid-state statistical mechanics of infinite systems but protein chains are finite systems, to remedy the conflict, the authors assumed that $N_{i,j}^{\mathrm{REF,p}}(r)$ increases in r^{α} with a to-be-determined constant α. Supposing that $\bar{u}_{i,j}(r) = 0$ for $r \geq r_{\mathrm{cut}}$ and $N_{i,j}^{\mathrm{REF,p}}(r_{\mathrm{cut}}) = N_{i,j}^{\mathrm{OBS,p}}(r_{\mathrm{cut}})$, $N_{i,j}^{\mathrm{REF}}(r)$ can be written as:

$$N_{i,j}^{\mathrm{REF}}(r) = \left(\frac{r}{r_{\mathrm{cut}}}\right)^{\alpha} \frac{\Delta r}{\Delta r_{\mathrm{cut}}} N_{i,j}^{\mathrm{OBS}}(r_{\mathrm{cut}}) \quad (7)$$

where $N_{i,j}^{\mathrm{OBS}}(r_{\mathrm{cut}}) = \sum_p N_{i,j}^{\mathrm{OBS,p}}(r_{\mathrm{cut}})$, and the summation is over all protein structures in the dataset. In Zhou and Zhou's training, α was set to 1.57 and r_{cut} to 14.5 Å.

Spherical non-interacting reference state (Dope-REF)

The Dope potential developed by Shen and Sali used a spherical noninteracting reference state,[17] which considered a sphere with a uniform uncorrelated atom density:

$$f^{\mathrm{REF,p}}(r,a) = \begin{cases} \frac{3r^2(r-2a)^2(r+4a)}{r_{\mathrm{cut}}^3\left(r_{\mathrm{cut}}^3 - 18a^2 r_{\mathrm{cut}} + 32a^3\right)} & r_{\mathrm{cut}} \leq 2a \\ \frac{6r^2(r-2a)^2(r+4a)}{16a^6} & r_{\mathrm{cut}} > 2a \end{cases} \quad (8)$$

where a is the size of the experimental structure sample p. Although protein structure is usually not a sphere, the size a can be defined as the radius of an effective sphere which has the same radius of gyration R_{g} as the sampled experimental structure, i.e. $a = \sqrt{5/3}R_{\mathrm{g}}$. We can thus calculate the potential by:

$$\bar{u}_{i,j}(r) = -RT \ln\left[\sum_p w_p \frac{N_{i,j}^{\mathrm{OBS,p}}(r)}{f^{\mathrm{REF,p}}(r,a) N_{i,j}^{\mathrm{OBS,p}}}\right] \quad (9)$$

where $N_{i,j}^{\mathrm{OBS,p}} = \sum_r^{r_{\mathrm{cut}}} N_{i,j}^{\mathrm{OBS,p}}(r)$, and w_p is the weight of the sampled experimental structure p which is calculated as the ratio between the number of atom pairs in this structure and the number of atom pairs in all sampled experimental structures, irrespective of the pair type. In some extent, the spherical noninteracting reference state can be regarded as an extended version of finite ideal-gas reference state with more theoretical details.

Atom-shuffled reference state (SRS-REF)

Unlike the above reference states which are either based on the sampled experimental structures or derived from certain analytical assumption, in the atom-shuffled reference state, all atomic positions were preserved while atom identities were shuffled within each of the experimental structures. $f_{i,j}^{\mathrm{REF}}(r)$ can be calculated from these shuffled structures, and we wrote it as $f_{i,j}^{\mathrm{shuffled}}(r)$.

$$\bar{u}_{i,j}(r) = -RT \ln\left[\frac{f_{i,j}^{\mathrm{OBS}}(r)}{f_{i,j}^{\mathrm{shuffled}}(r)}\right] \approx -RT \ln\left[\frac{N_{i,j}^{\mathrm{OBS}}(r)}{N_{i,j}^{\mathrm{shuffled}}(r)}\right] \quad (10)$$

The HA_SRS potential developed by Rykunov and Fiser used this reference state.[16] The authors presented

three shuffle patterns including residue-shuffled, sequence-shuffled, and atom-shuffled. Here, we implemented the last one. The dataset used to generate the shuffled structures is the same as that used to calculate $N_{i,j}^{\mathrm{OBS}}(r)$. We shuffled every experimental structure more than one million times by randomly exchanging the identity of two atoms.

Random-walk chain reference state (RW-REF)

Since the starting point of protein folding is the amino acid sequence, the RW potential developed by Zhang and Zhang used an ideal random-walk (RW) chain of a rigid step length as the reference state.[14] This RW model mimics well the generic entropic elasticity and inherent connectivity of polymer protein molecules and yet ignores the atomic interactions of amino acids. According to the polymer theory in the freely-jointed chain model, the reference probability can be written as:

$$f^{\mathrm{REF,p}}(r) = \int f^{\mathrm{REF,p}}(r,n)\mathrm{d}n = \sum_{n=1}^{N} 4\pi r^2 \left(\frac{3}{2\pi n l^2}\right)^{3/2} \exp\left(-\frac{3r^2}{2nl^2}\right)\Delta r \quad (11)$$

where N is the number of residues in the sample protein p, and l is the Kohn length. As is done in finite ideal-gas reference state, given a cutoff distance and assuming $N_{i,j}^{\mathrm{REF,p}}(r_{\mathrm{cut}}) = N_{i,j}^{\mathrm{OBS,p}}(r_{\mathrm{cut}})$, we can get:

$$N_{i,j}^{\mathrm{REF}}(r) = \sum_{p} \left(\frac{r}{r_{\mathrm{cut}}}\right)^2 \frac{\sum_{n=1}^{N} \exp(-3r^2/2nl^2)/n^{3/2}}{\sum_{n=1}^{N} \exp(-3r_{\mathrm{cut}}^2/2nl^2)/n^{3/2}} N_{i,j}^{\mathrm{OBS,p}}(r_{\mathrm{cut}}) \quad (12)$$

The value of l^2 was set to 460 in the RW potential, under which the potential had the best performance.[14]

RESULTS

We constructed the six potentials based on the same dataset of 1022 protein structures using the reference models as formulated in Eqs. (4)–(12). Our evaluations are focused on the ability of prioritization of the native and near-native structures, as well as the energy-TM-score correlations. To establish the generality of the analysis, we apply the potentials to various decoy sets generated from different methods.

CASP decoy set

First, we evaluate the potentials in the structural models generated in CASP5-CASP8 experiments as collected by Rykunov and Fiser,[13] which include 143 targets and 2628 models. Since these structural models were predicted blindly by all CASP participants using the state-of-the-art methods, this set represents the most diverse decoys and the selection of best decoy models has practical use.

Table I
Performance of Six Potentials in CASP Decoys

Potential[a]	N_{nat}[b]	$\mathrm{Rank}_{\mathrm{nat}}$[c]	Z-score[d]	R/TM[e]	Corr[f]
RAPDF-REF	90/143	2.0/19.4	1.46	10.88/0.581	−0.46
KBP-REF	107/143	1.6/19.4	1.65	7.60/0.613	−0.63
Dfire-REF	80/143	2.9/19.4	1.23	6.60/0.644	−0.57
Dope-REF	93/143	1.9/19.4	1.50	10.71/0.584	−0.46
SRS-REF	92/143	2.1/19.4	1.44	7.49/0.607	−0.49
RW-REF	79/143	3.2/19.4	1.19	6.56/0.646	−0.56

[a]Potentials that we reconstructed from a unified structure dataset by using corresponding reference state models from Eqs. (4)–(12).
[b]The number of targets with the native structure ranked as first versus the total number of test proteins.
[c]The average rank of the native structures versus the average number of conformations per target.
[d]Z-score = $(E_{\mathrm{average}} - E_{\mathrm{native}})/\sigma$, where E_{native} is the energy of the native structure, and E_{average} is the average energy of all decoys. σ is the energy deviation of all decoys.
[e]The average RMSD and TM-score to the native of the first ranked models.
[f]The average Pearson correlation between energy and TM-score of decoys.

Table I summarizes the performance results of all six potentials on prioritizing the CASP models. Here and after, we denote the potential based on certain reference as "xxx-REF." We can see that KBP-REF outperforms all other potentials on most evaluation criteria except for the average RMSD and TM-score of the first ranked models which are slightly lower than Dfire-REF and RW-REF. RAPDF-REF, Dope-REF, and SRS-REF have similar performances, and select about 10 more native structures than Dfire-REF and RW-REF; however, these potentials have generally lower correlations than Dfire-REF and RW-REF. The performances of Dfire-REF and RW-REF are similar which have the best average RMSD and TM-score of the first ranked models.

Take T0233 as a typical example, the correlations from different potentials are varied (Fig. 1). Dfire-REF and RW-REF fail to select the native structure while their correlation coefficients are relatively better, which demonstrates the potential usefulness of the potentials to guide the folding simulations. In Supporting Information Figures S1–S3, we show three additional examples from T0137, T0211, and T0423, which have three level of high, medium, and low potential-TM-score correlations, respectively. They have a similar tendency in the energy-TM-score correlations as what we have seen in Figure 1 and Table I.

Ig_structal_hires decoy set

Next we applied the potentials to three target decoy sets from the Decoys 'R' Us,[35] including *ig_structal_hires*, *fisa_casp3*, and *lattice_ssfit*. The *ig_structal-_hires* decoy set contains 20 immunoglobulin proteins and the decoy structures were built by comparative modeling program

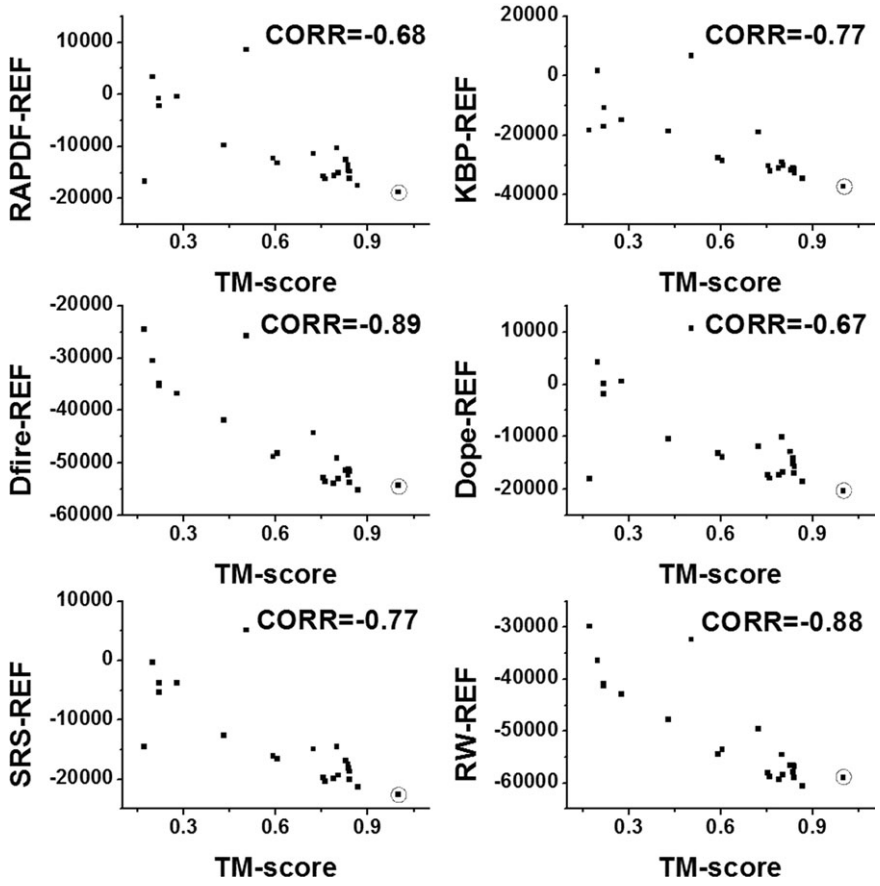

Figure 1

A typical example of energy-TM-score correlation from T0233 in the CASP decoy set, where the energy of each decoy conformation is calculated by six different potentials. The native structure is highlighted by the open circles.

segmod[36] using other immunoglobulins as templates. As shown in Table II, RAPDF-REF performs the best on selecting native structures, while KBP-REF has the highest energy-TM-score correlation with a typical example shown in Figure 2. The average RMSD and TM-score of the first ranked models from RW-REF is slightly better than other potentials. In Supporting Information Figures S4–S6, we present three additional examples of this decoy set with the decoy structures from 1mfa, 1vge, and 7fab, respectively.

Fisa_casp3 decoy set

There are five decoy sets in *fisa_casp3*, and each set contains about 1400 decoy conformations. The backbone conformations of these decoys were generated by Rosetta program [21] which assembled the models using fragments of other solved protein structures; side-chain atoms were then added by SCWRL.[37] Since the decoy conformations were from *ab initio* modeling, most structures have a low TM-score (<0.5). In this low-resolution region, all potentials have an almost negligible correlation with the TM-score. Figure 3 shows four proteins by RW-REF, where the energy-TM-score correlation coefficient is below 0.4 for all protein targets. A similar tendency is seen in all other potentials on this decoy set (see Supporting Information Figs. S7–S11).

Probably because the decoys are mainly distributed at low TM-score (far from the native), the native structures in this set are relatively easy to recognize by most potentials. As shown in Table III, all potentials, except for KBP-REF, can correctly recognize the native in four of five targets. The remaining target is from 1b0nB whose

Table II

Performance of Potentials in *ig_structal_hires* of Decoys 'R' Us[a]

Potential	N_{nat}	$Rank_{nat}$	*Z*-score	R/TM	Corr
RAPDF-REF	11/20	5.4/20.0	1.05	2.21/0.945	−0.77
KBP-REF	6/20	6.1/20.0	0.69	2.16/0.945	−0.86
Dfire-REF	2/20	11.1/20.0	0.15	2.14/0.948	−0.81
Dope-REF	10/20	6.3/20.0	0.82	2.32/0.945	−0.80
SRS-REF	10/20	5.8/20.0	0.94	2.20/0.946	−0.79
RW-REF	1/20	11.9/20.0	-0.04	2.11/0.949	−0.80

[a]Notations are the same as that in Table I.

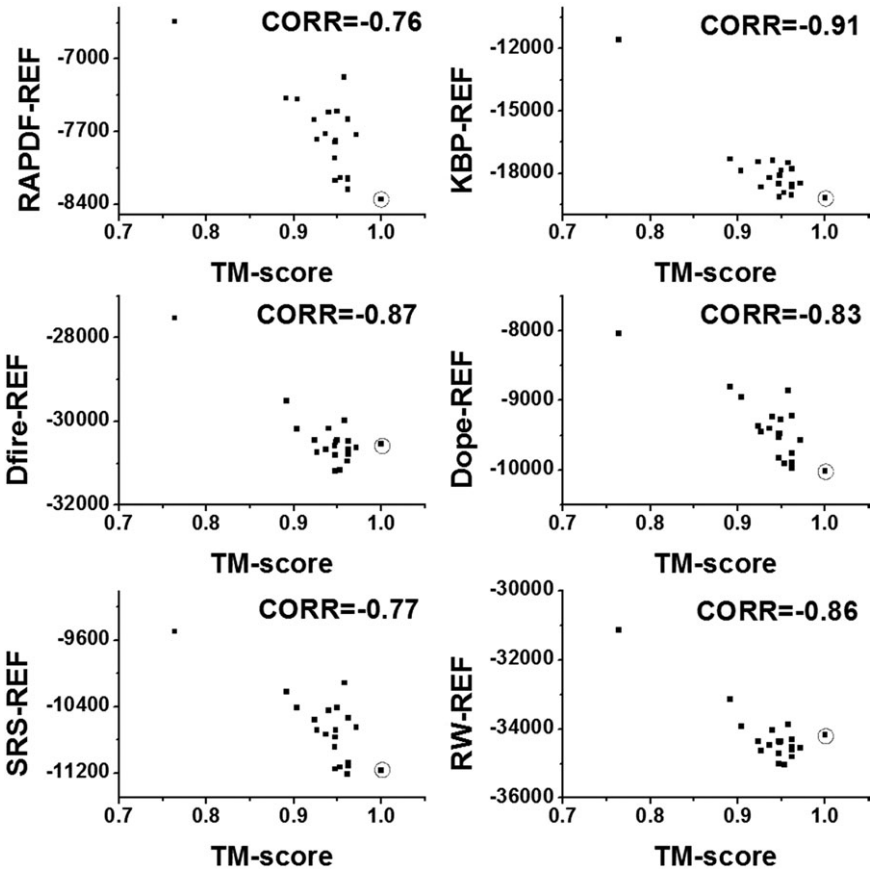

Figure 2

A typical example of energy-TM-score correlation from 1fgv in *ig_structal_hires* of Decoys "R" Us, where the energy of each decoy conformation is calculated by six different potentials. The native structure is highlighted by the open circles.

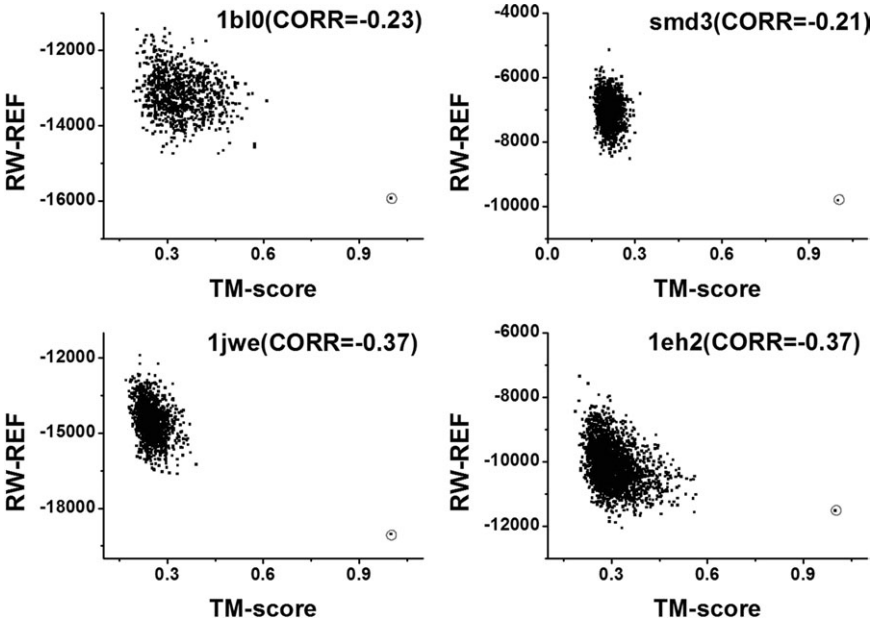

Figure 3

Examples of energy-TM-score correlation by RW-REF in *fisa_casp3* of Decoys "R" Us. The native structure is highlighted by the open circles.

Table III
Performance of Six Potentials in *fisa_casp3* of Decoys 'R' Us[a]

Potential	N_{nat}	$Rank_{nat}$	*Z*-score	R/TM	Corr
RAPDF-REF	4/5	203.8/1439.0	3.45	10.97/0.299	−0.11
KBP-REF	2/5	108.4/1439.0	2.20	11.99/0.294	−0.17
Dfire-REF	4/5	7.6/1439.0	4.62	11.00/0.298	−0.26
Dope-REF	4/5	104.8/1439.0	3.88	11.46/0.265	−0.12
SRS-REF	4/5	133.6/1439.0	3.98	10.97/0.299	−0.14
RW-REF	4/5	4.8/1439.0	4.78	10.70/0.310	−0.28

[a]Notations are the same as that in Table I.

Table IV
Performance of Six Potentials in *lattice_ssfit* of Decoys 'R' Us[a]

Potential	N_{nat}	$Rank_{nat}$	*Z*-score	R/TM	Corr
RAPDF-REF	8/8	1/1999.6	5.30	10.42/0.241	−0.07
KBP-REF	8/8	1/1999.6	5.53	10.98/0.237	−0.12
Dfire-REF	8/8	1/1999.6	7.65	9.77/0.248	−0.15
Dope-REF	8/8	1/1999.6	5.39	10.00/0.245	−0.08
SRS-REF	8/8	1/1999.6	5.40	10.52/0.243	−0.08
RW-REF	8/8	1/1999.6	8.30	10.08/0.250	−0.17

[a]Notations are the same as that in Table I.

native structure has an irregular topology of the extended two-helix bundle which is stabled only when intertwined with the Chain A of the protein. All potentials, ranking on the isolated domain without counting the interaction with Chain A, failed to recognize the native state. The overall ranking and correlation results of *fisa_casp3* are listed in Table III, where the RW-REF performs relatively better than other potentials on every aspect.

Lattice_ssfit decoy set

The *lattice_ssfit* decoy set contains eight small proteins generated by *ab initio* enumerations of possible conformations in a lattice system.[38] Similar to the *fisa_casp3*, most of the decoy structures have a low TM-score. Thus, the recognition of the native structure is relatively easy and all potentials could recognize the native state of all targets with a high *Z*-score. Accordingly, there is almost no correlation between energy and TM-score as shown in Figure 4, which was based on RW-REF that has the highest average correlation coefficient. In Supporting Information Figures S12–S16, we present examples from other five potentials on the same set of proteins, where a similar correlation range is seen in these potentials. Again, as shown in Table IV, RW-REF outperforms all potentials in all the criteria in this decoy set.

MOULDER decoy set

We also tested the potentials in the MOULDER decoy sets which were generated by the comparative modeling program MODELLER where close homologous templates have been used to guide the model generations.[39] To cover a wider RMSD range, we have selected templates with alignments ranging from 0 to 100% of the native overlaps. As shown in Table V, all six potentials can easily select the native structures for the majority of targets with an appreciable *Z*-score. The averages of the energy-TM-score correlation also reach to a high level with coefficient >0.75 for all potentials.

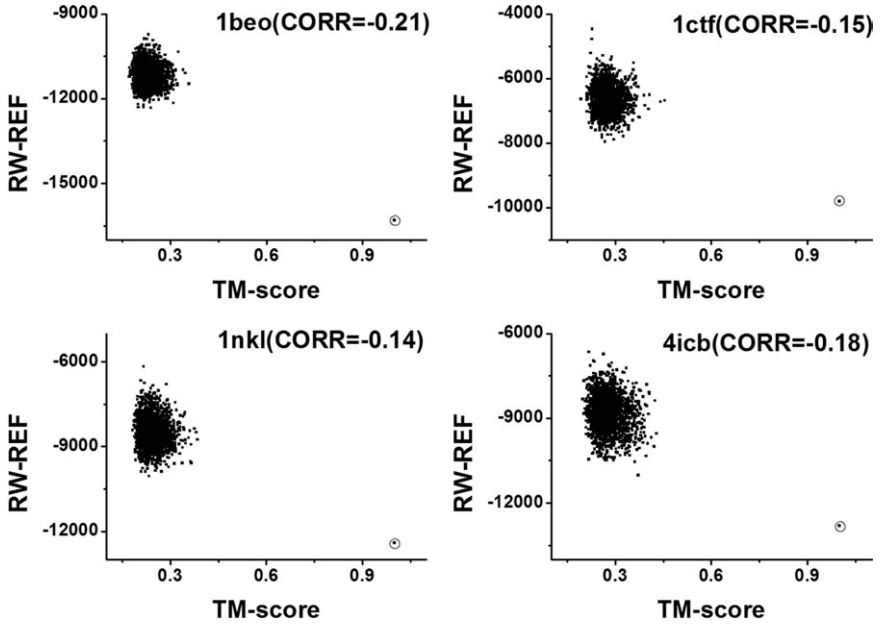

Figure 4
Examples of energy-TM-score correlation by RW-REF in *lattice_ssfit* of Decoys "R" Us. The native structure is highlighted by the open circles.

Table V
Performance of Six Potentials in the MOULDER Decoy Sets[a]

Potential	N_{nat}	$Rank_{nat}$	*Z*-score	R/TM	Corr
RAPDF-REF	19/20	5.3/301.0	3.05	4.60/0.746	−0.78
KBP-REF	19/20	2.4/301.0	2.42	4.57/0.750	−0.87
Dfire-REF	19/20	6.0/301.0	2.98	3.98/0.771	−0.88
Dope-REF	19/20	5.3/301.0	3.23	4.34/0.761	−0.79
SRS-REF	19/20	4.3/301.0	3.18	4.32/0.750	−0.81
RW-REF	19/20	4.5/301.0	2.94	4.45/0.752	−0.88

[a]Notations are the same as that in Table I.

This high correlation value is partly due to the wider range of the decoy distributions because by definition the correlation coefficient can achieve a higher value in the wider distributed decoys than the narrow distributed ones even with a similar level of decoy fluctuations. Second, the decoy structures in MOULDER were generated by comparative modeling which keeps most of the template structural unchanged. These are different from the decoys generated by *ab initio* folding that have all structure regions reassembled from scratch. Thus, the statistical potentials, which are all developed from the PDB structure library, may tend to have a better discrimination power on the homology-based decoys due to some level of memory effects.

Among all the potentials, Dfire-REF has a relatively stronger energy-TM-score correlation and recognition accuracy for near-native structures according Table V, but its performance to recognize the native structures is slightly worse than other potentials. Figure 5 shows four typical examples by Dfire-REF. Indeed, the decoys have a quite uniformed distribution spanning a much larger range than the *ab initio* folding decoys. The correlation is consequently higher than that in other decoy sets. The illustrated examples for other five potentials are shown in Supporting Information Figures S17–S21.

I-TASSER decoy set-II

Finally, we used the I-TASSER Decoy Set-II which has the coarse-grained models first generated by iterative Monte Carlo fragment assembly and then refined by GROMACS4.0 MD simulation.[14] This set represents a typical procedure of protein structure predictions combining template-based modeling and atomic-level structure refinements. As shown in Table VI, the six potentials can select the majority of native structures with discrepancies less than 9. RW-REF outperforms others on all criteria, and Dfire-REF takes second place. The gap between the best and worst performing potentials on energy-TM-score correlation is as high as twenty percent.

Figure 6 presents four typical examples of I-TASSER Decoy Set-II by RW-REF. The decoy conformations from 1abv_, 1gjxA, and 1vcc_ have low TM-score, which have accordingly a low energy-TM-score correlation value. However, in decoy set of 1thx_, the decoy conformations gather into two clusters, one cluster is with TM-score around 0.8 and the other is with TM-score around 0.5. The correlation value for this target is much stronger

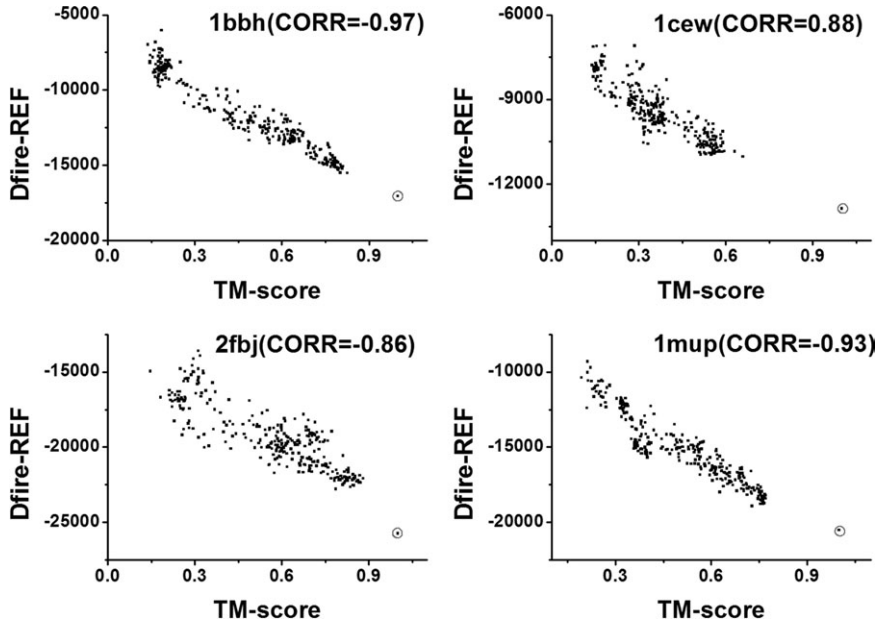

Figure 5
Examples of energy-TM-score correlation by Dfire-REF in the MOULDER decoy sets. The native structure is highlighted by the open circles.

Table VI
Performance of Six Potentials in I-TASSER Decoy Set-II[a]

Potential	N_{nat}	$Rank_{nat}$	*Z*-score	R/TM	Corr
RAPDF-REF	49/56	23.48/441.2	5.28	6.20/0.545	−0.34
KBP-REF	45/56	34.43/441.2	3.82	5.38/0.549	−0.42
Dfire-REF	53/56	6.79/441.2	5.08	5.23/0.561	−0.52
Dope-REF	50/56	18.54/441.2	5.43	6.12/0.548	−0.35
SRS-REF	49/56	25.68/441.2	5.11	5.72/0.552	−0.38
RW-REF	53/56	2.48/441.2	5.45	5.14/0.568	−0.54

[a]Notations are the same as that in Table I.

(−0.88). This data demonstrates again that as a necessary condition, the decoys should cover a broad range of resolution in order to have a high apparent value of correlation coefficient (see Fig. 5). In Supporting Information Figures S22–S26, we present the examples of other five potentials on the same set of proteins, where they all have a higher correlation coefficient on 1thx_.

DISCUSSION

The importance of reference state

The characteristics of native conformations can clearly show up only when comparing with nonnative ones, where the nonnative conformations serve as the reference state. Our brains can subconsciously set a reference state for every judgment or evaluation with powerful inertia and intelligence, while computer-based statistical potentials cannot do so. We must design a reference state in advance and integrate it into the formula of potential. The usefulness of statistical potential depends on its ability to distinguish native conformations or find best models from nonnative conformations. So the key task for the potential construction is to explore and utilize the structural differences between native and nonnative conformations.[9] As to atomic distance-dependent pair-wise contact potential, what we concern are the differences of atom pair distribution between native and nonnative conformations. The distribution of native conformations can be obtained through statistics on the PDB library. The problem is how to get the distribution of nonnative conformations, or in other words, how to describe the reference state. Any reference state can only cover a specific conformation space, thus the potential should better be applied to the structures that the reference state can suitably cover. The diverse performances of the potentials on different reference states imply that the potentials are strongly shaped by its reference state.

Statistical reference state versus theoretical reference state

In the six reference states considered here, the averaging, quasi-chemical approximation and atom-shuffled reference states are primarily base on statistics of experimental structures.[14] The statistical samples for averaging and quasi-chemical approximation reference states are directly from experimental protein structures, and atom-shuffled reference state uses a set of shuffled experimental

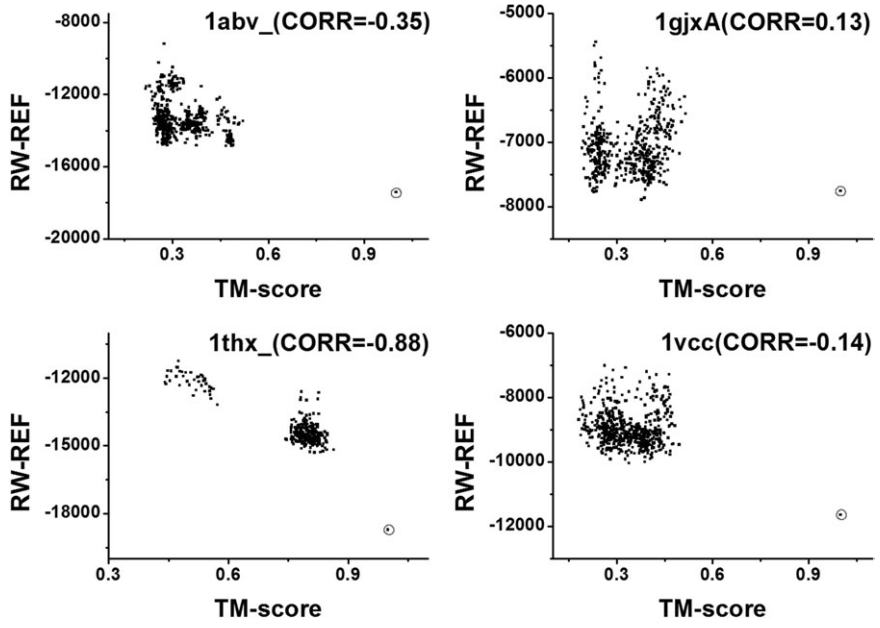

Figure 6
Examples of energy-TM-score correlation by RW-REF in I-TSAAER Decoy Set-II. The native structure is highlighted by the open circles.

conformations. Since there is no proper nonnative dataset exploited, the reference state derived from the native protein structures may not appropriately reflect the conformational sampling of nonnative states encountered in real folding simulations. On the contrary, the finite ideal-gas, spherical non-interacting and random-walk chain reference states are from theoretical reference state for they are mainly based on theoretical assumptions and effectively circumvent concrete statistical processes, which are often oversimplified for real modeling procedure. In this context, a reference state considering the statistics of realistic computational simulation decoys is probably essential.

The universality and pertinence of statistical potential

The results presented here show that no potential can always outperform others in different decoy sets. Even in the same decoy set they often rank inconsistently in different evaluation criteria. As described earlier, the distinction among the six potentials merely reflects in their different reference states, from which their diverse performances consequently arise. No matter how to deal with the reference state, the conformation spaces that different reference states can cover are different. For example, the averaging reference state was based on native structures and can be a suitable representation of near-native conformations; while the finite ideal-gas reference state is based on the assumption of finite ideal-gas and thus can roughly cover a broad conformation space. But what method is the more suitable? If we want the potential to be efficient under a broader application environment, namely that the universality of potential is emphasized, we should calculate the reference state basing on a more general conformation space. However, the pertinence of potential would be compromised while enhancing its universality, and too much emphasis on universality is likely to make the potential perform poorly in any application environment. As for the six reference states we used here, the conformation spaces they can cover are obviously different, which consequently makes the potentials based on them have respective universality and pertinence. It is the distinction on universality and pertinence that makes the potentials perform diversely in different decoy sets. To further enhance the performance of statistical potential, we can envisage the range of application at the beginning of potential construction while not being keen on its universal validity, and calculate reference state based on the specific application environment. For example, if the potential is designed mainly for assessing and refining the conformations produced by certain prediction method, we should probably take a nonredundant conformation set produced by this method as the statistical samples of reference state, and both expanded and narrowed conformation space of the sample structures would have negative impact on its performance.

Calculation procedure of statistical potential

There are two ways that we can choose in the calculation procedure of statistical potential. One is to divide the observed contact numbers in the entire sample dataset by the referential contact numbers first and then take its negative logarithm; the other is to divide the observed contact numbers in a single sample protein by corresponding referential contact numbers and then combine the results over the entire dataset, and finally take its negative logarithm. While the observed contact numbers in a single sample protein would likely be too sparse to allow an effective statistics.[12] We tested the above two ways in the calculation procedure of Dope-REF and RW-REF potentials since both of their reference states are related to the protein size. The result shows that the Dope-REF potentials calculated in two ways perform similarly, but the RW-REF potential calculated in the first way performs much better than that calculated in the second way (detail data not shown). With a view to the conventional calculation procedures related to different reference states, in this paper we used the first way to calculate all the potentials except for the Dope-REF potential, which is calculated in the second way.

Effects of TM-score (or RMSD) distribution of decoy set to evaluation criteria

All decoy sets we used here include the native structures. There are often large gaps on TM-score between the native structures and their decoy conformations, which may partly make the native structure selection much easier than the discrimination of decoys in different accuracy. As shown in the previous section, the criteria related to the native structure selection (N_{nat}, $Rank_{nat}$, and Z-score) generally get better values than those related to the discrimination of decoys in different accuracy (R/TM and CORR). We investigated into the TM-score and RMSD distributions of decoy sets and found there are large discrepancies among different decoy sets. When the distributions are narrow and concentrated, R/TM and CORR might be poor. For instance, energy-TM-score correlation calculated in decoy set 1thx_ is much better than that calculated in the other three set in Figure 6, which is clearly linked to the particular TM-score distribution of decoy set 1thx_. Overall, these data indicated that the potentials are merely able to distinguish the decoys in a coarse level and their discriminatory powers remain to be enhanced.

Here, it is important to note that our assessment criteria are more practices-oriented rather than physics-based, although it is important to have the correct reference state that is as close as possible to physics. One reason is that most of the reference states are based on some as-

pect of physical rules in their original developments, but we do not have an objective criterion to quantitatively assess how close the potentials are to physics. The energy-TM-score correlations and the Z-score of the native structures over decoys, on the other hand, can give a quantitative assessment of the potentials in their ability of assisting protein folding and decoy recognition. These criteria have been widely used in the development and assessment of various statistical potentials.[9,12,14–15,17] Second, due to the limit size of the current structural databases, the "physically correct" reference states do not always work the best in practical uses. Although the ideal potential should be both physically and practically sound, here we prefer to choose those that can have best performance in practical applications, when a compromise has to make between them and especially when we do not have a clear criterion to assess the physical correctness of the potentials.

CONCLUSION

Starting with different reference states, we constructed six atomic distance-dependent pair-wise contact potentials based on a uniform sampling dataset and bin-width procedure. These potentials were assessed by virtue of six independent decoy sets. Overall, the random-walk chain model outperformed others in three sets of decoy sets, while reference states based on ideal-gas, quasi-chemical approximation and averaging sample did so in one decoy set separately. Nevertheless, the performance of the potentials fluctuated depending on the decoy sets. No potential could dominate the structural selection and energy-TM-score correlation in all the cases. Our analyses demonstrate that statistical potential has its universality and pertinence which is decided by the reference state and the decoy sets. The optimal reference state should probably be derived by the consideration of the conformational sampling of specific modeling simulations.

The somewhat contradictory assessment results and especially the performance dependence on decoy distributions indicate that the current mean-force statistical potential developments are far from the true solution (if it exists at all). This result is consistent with the well-established agreement in the community that the single-model based quality assessment method cannot compete with the consensus-based approaches in near-native structure recognitions.[40–44] However, the performance of statistical potentials is still significantly better than the random model selections based on our unpublished data. Recent studies showed that a combination of the single-model potentials with structural clustering can outperform that based on consensus,[45,46] which may represent another promising avenue to the improvement of the single-model statistical potentials.

REFERENCES

1. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins Struct Funct Genet 1995;21:167–195.
2. Skolnick J. In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol 2006;16:166–171.
3. Floudas CA. Computational methods in protein structure prediction. Biotechnol Bioeng 2007;97:207–213.
4. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
5. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
6. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Computer-aided Mol Des 1993;7:473–501.
7. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676–688.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
9. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction1. J Mol Biol 1998;275:895–916.
10. Sippl MJ. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
11. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins1. J Mol Biol 1997;267:707–726.
12. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins: Struct Funct Bioinformatics 2001;44:223–232.
13. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics 2010;11:128.
14. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One 2010;5:e15386.
15. 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.
16. Rykunov D, Fiser A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins 2007;67:559–568.
17. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
18. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins: Struct Funct Bioinform 1993;16:92–112.
19. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins: Struct Funct Bioinform 1999;36:357–369.
20. Xu Y, Xu D. Protein threading using PROSPECT: design and evaluation. Proteins 2000;40:343–354.
21. 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.
22. Bowie JU, Luthy R, Eisenberg D. A Method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
23. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.

24. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
25. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.
26. Floudas CA, Fung HK, McAllister SR, Monnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem Eng Sci 2006;61:966–988.
27. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci USA 2007;104: 3177–3182.
28. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19:1784–1795.
29. Gatchell DW, Dennis S, Vajda S. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins Struct Funct Genet 2000;41:518–534.
30. Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence1. J Mol Biol 1997;272:276–290.
31. Wallqvist A, Jernigan R, Covell D. A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. Protein Sci 1995;4:1881–1903.
32. 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.
33. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
34. Wang G, Dunbrack JrRL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
35. Samudrala R, Levitt M. Decoys 'R'Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
36. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992;226:507–533.
37. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
38. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Citeseer 1999;505–506.
39. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
40. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
41. Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D. MUFOLD: a new solution for protein 3D structure prediction. Proteins 2010;78:1137–1152.
42. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69:184–193.
43. Wang Z, Eickholt J, Cheng J. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 2011;27: 1715–1716.
44. McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics 2007;8:345.
45. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 2009;77:100–113.
46. Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 2011;79:147–160.