Yu *et al. BMC Bioinformatics* (2019) 20:299
https://doi.org/10.1186/s12859-019-2898-y

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

ANDIS: an atomic angle- and distance-dependent statistical potential for protein structure quality assessment

Zhongwang Yu[1], Yuangen Yao[1], Haiyou Deng[1,2*] and Ming Yi[1,2*]

Abstract

Background: The knowledge-based statistical potential has been widely used in protein structure modeling and model quality assessment. They are commonly evaluated based on their abilities of native recognition as well as decoy discrimination. However, these two aspects are found to be mutually exclusive in many statistical potentials.

Results: We developed an atomic ANgle- and DIStance-dependent (ANDIS) statistical potential for protein structure quality assessment with distance cutoff being a tunable parameter. When distance cutoff is ≤9.0 Å, "effective atomic interaction" is employed to enhance the ability of native recognition. For a distance cutoff of ≥10 Å, the distance-dependent atom-pair potential with random-walk reference state is combined to strengthen the ability of decoy discrimination. Benchmark tests on 632 structural decoy sets from diverse sources demonstrate that ANDIS outperforms other state-of-the-art potentials in both native recognition and decoy discrimination.

Conclusions: Distance cutoff is a crucial parameter for distance-dependent statistical potentials. A lower distance cutoff is better for native recognition, while a higher one is favorable for decoy discrimination. The ANDIS potential is freely available as a standalone application at http://qbp.hzau.edu.cn/ANDIS/.

Keywords: Statistical potential, Pair-wise interaction, Protein decoy set, Distance cutoff, Protein structure prediction

Background

The primary mission in protein structure prediction is to develop accurate energy functions for conformational search [1–5], model refinement [6–9], and model quality assessment [10–12]. However, because of the big size, the flexibility and the presence of solvent molecules, proteins are still extremely difficult to model with physics-based potential [13, 14]. especially when quantum mechanical calculation is involved [15]. The knowledge-based potential [16–19], which is extracted from the experimental structures deposited in Protein Data Bank, has been playing an increasingly important role in protein structure prediction since its emergence in 1990s [20–22]. Varieties of structural features were used to derive knowledge-based potentials, such as residue solvent accessibility [23, 24], residue or atom contact [25, 26], atom-pair distance distribution [27–29], side-chain orientation [16, 30, 31] and so on. The Boltzmann law and probability theory are commonly employed to convert the observed frequencies of specific structural features into statistical potentials [17, 20].

To evaluate a potential function, basically the following two aspects need to be considered: (a) can the potential recognize native or near-native structure from non-native structures? (b) can the energy scores given by the potential well reflect the structural qualities of different prediction models? Both aspects can be assessed by applying the potential to various protein structure decoy sets [32–35]. In fact, the majority of statistical potentials were derived by optimizing both

* Correspondence: hydeng@mail.hzau.edu.cn; yiming@mail.hzau.edu.cn
[1]Department of Physics, College of Science, Huazhong Agricultural University, Wuhan 430070, China
Full list of author information is available at the end of the article

performances in native recognition and decoy discrimination [30, 36–38]. However, native recognition emphasizes the differences of overall structure quality between native and decoy structures (e.g., by maximizing the all-atom energy difference between the native structure and other non-native structures). While decoy discrimination generally focuses on the backbone differences among decoy structures (e.g., by enhancing the correlation of potential score with GDT_TS, TM-score etc.). They are actually in different levels (atomic and residual levels, respectively), thus the coupling of them would require a trade-off in potential optimization. Our previous work clearly indicates that the potential's abilities of native recognition and decoy discrimination cannot be optimized simultaneously with the same parameter sets [39]. For protein structure modeling, the ability of decoy discrimination is more crucial. Commonly the energy function targeted to the modeling method is used. But for researchers who want to choose a better structure for biological analysis, the overall structure quality with native structure as the gold standard should be emphasized.

In this work, we developed an atomic angle- and distance-dependent (ANDIS) statistical potential for protein structure quality assessment. A total of 167 residue-specific, heavy atom types are considered. As done in GOAP potential [37], we define a local coordinate system for every heavy atom in protein structure based on the positions of the atom and two of its bonded neighboring atoms. The pair-wise interaction between atoms with distance < 15.0 Å and residue separation ≥7 are considered. 5 angles (4 polar angles and 1 dihedral angle) are calculated according to the relative orientation of local coordinate systems between the two interacted atoms. Since the angles are strongly associated with side-chain packing and hydrogen-bonding, the ANDIS potential naturally integrates the atomic distance-dependent and orientation-dependent interactions. The distance cutoff is designed to be adjustable from 7 Å to 15.0 Å. A lower distance cutoff (< 9.5 Å) is recommended for native recognition, and the energy of each atom-pair with distance below 9.5 Å is weighted based on the degree of mutual exposure. On the contrary, a higher distance cutoff (≥10 Å) is recommended for decoy discrimination, and a distance-dependent atom-pair potential with random-walk reference state [30] is combined with the angle energies to enhance the ability of decoy discrimination.

We benchmarked ANDIS with a comprehensive list of publicly available statistical potentials (Dfire [36], RW [30], GOAP [37], DOOP [40], etc.), via 632 protein structural decoy sets collected from diverse sources. The results indicate that ANDIS significantly outperforms other reported statistical potentials in terms of native structure recognition. The effects of different protein datasets and distance cutoffs on ANDIS's performance are also comprehensively investigated. A detailed discussion is given below.

Methods

Experimental protein structures for calculating the potentials

A non-redundant structural dataset of 3519 protein chains were used for potential derivation. It was culled by PISCES [41] from Protein Data Bank with pairwise sequence identity < 20%, resolution < 2.0 Å and R-factor < 0.25 (only the structures determined by X-ray crystallography were considered). The original list from PISCES contains about 7000 protein chains. We excluded the proteins with incomplete, missing or non-standard residues and the proteins with length < 30 or > 1000 residues. The dataset is publicly available at http://qbp.hzau.edu.cn/ANDIS/.

Definition of distance-dependent angles

Various aspects of structural features (e.g., solvent accessibility, electrostatic interaction, contact, distance, torsional angle) can be used to derive statistical potential, with distance-dependent pair-wise interaction being the most commonly adopted. In ANDIS potential the atom-pairs with residue separation (in protein sequence) ≥ 7 and distance < 15.0 Å are considered. There are a total of 167 residue-specific, heavy (non-hydrogen) atom types in the 20 common amino acids. The distance between atom pair is divided into 29 bins (first bin is 0–2.2 Å, bin wide is 0.4 Å from 2.2 Å to 7.0 Å and 0.5 Å from 7.0 Å to 15.0 Å). ANDIS is designed to capture the structural characteristics embedded in the relative orientation of interacting atoms as well as in the distance distribution of atom-pairs.

As shown in Fig. 1, a local coordinate system is established for each atom based on itself and 2 neighboring bonded atoms (the next-neighbor, bonded atom is used if there is only one bonded heavy atom). To specify the relative orientation of the two coordinate systems, 5 distance-dependent angles are defined, including 4 polar angles (θ_a, φ_a, θ_b, φ_b for the orientation of $\boldsymbol{r}_{ab}$ or $\boldsymbol{r}_{ba}$ in the local coordinate system) and 1 dihedral angle (χ between plane $\boldsymbol{r}_{ab} \times \boldsymbol{V}_z$ (a) and plane $\boldsymbol{V}_z$ $(b) \times \boldsymbol{r}_{ba}$). A more detailed description of these angles is given by Zhou and Skolnick for their GOAP potential. [37]

The values of θ_a, θ_b, φ_a, φ_b and χ are equally spitted into 12 bins. Thus the original size of the statistical

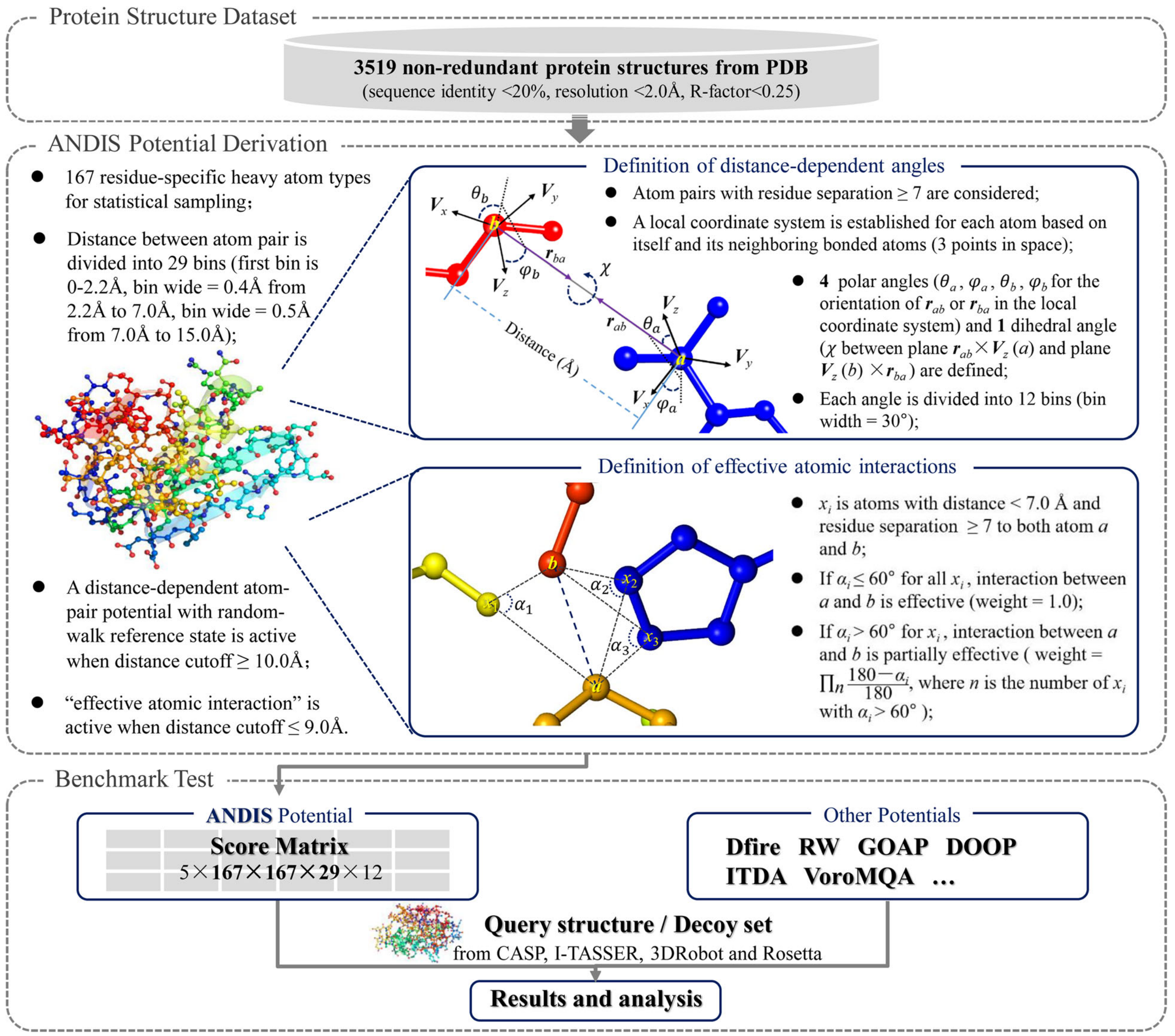
Protein Structure Dataset
3519 non-redundant protein structures from PDB
(sequence identity <20%, resolution <2.0Å, R-factor<0.25)
ANDIS Potential Derivation
● 167 residue-specific heavy atom types for statistical sampling;
● Distance between atom pair is divided into 29 bins (first bin is 0-2.2Å, bin wide = 0.4Å from 2.2Å to 7.0Å, bin wide = 0.5Å from 7.0Å to 15.0Å);
● A distance-dependent atom-pair potential with random-walk reference state is active when distance cutoff ≥ 10.0Å;
● "effective atomic interaction" is active when distance cutoff ≤ 9.0Å.
Definition of distance-dependent angles
● Atom pairs with residue separation ≥ 7 are considered;
● A local coordinate system is established for each atom based on itself and its neighboring bonded atoms (3 points in space);
● 4 polar angles (θ_a, φ_a, θ_b, φ_b for the orientation of r_{ab} or r_{ba} in the local coordinate system) and 1 dihedral angle (χ between plane $r_{ab} \times V_z(a)$ and plane $V_z(b) \times r_{ba}$) are defined;
● Each angle is divided into 12 bins (bin width = 30°);
V_x
θ_b
V_y
b
r_{ba}
φ_b
V_z
χ
Distance (Å)
r_{ab}
θ_a
V_z
a
V_y
V_x
φ_a
Definition of effective atomic interactions
● x_i is atoms with distance < 7.0 Å and residue separation ≥ 7 to both atom a and b;
● If $\alpha_i \le 60°$ for all x_i, interaction between a and b is effective (weight = 1.0);
● If $\alpha_i > 60°$ for x_i, interaction between a and b is partially effective (weight = $\prod_n \frac{180-\alpha_i}{180}$, where n is the number of x_i with $\alpha_i > 60°$);
b
x_1
α_1
α_2
x_2
x_3
α_3
a
Benchmark Test
ANDIS Potential
Score Matrix
5×167×167×29×12
Other Potentials
Dfire RW GOAP DOOP
ITDA VoroMQA …
Query structure / Decoy set
from CASP, I-TASSER, 3DRobot and Rosetta
Results and analysis

Fig. 1 The flowchart of our studies. Step 1. PDB dataset preparation; Step 2. Potential derivation; Step 3. Benchmark test

matrix is 5 × 167 × 167 × 29 × 12. In statistics, we ignored the angle distributions (e.g., the second distance bin 2.2 Å–2.6 Å of atom-pair CYS N – PHE CE2 for angle φ_a) whose occurrences were below 20 to ensure reasonable statistics.

Definition of effective atomic interactions

In order to capture the pair-wise interactions that are more likely to be physically relevant, we consider only the "effective atomic interactions" in our potential [42]. As shown in Fig. 1, the physical exposure between atom a and b is evaluated by calculating the angle α_i ($\angle a x_i b$) for every atom x_i with distance < 7.0 Å to both atom a and b. A large angle α_i means that atom a and b are shielded by atom x_i. Here we consider the interaction of atom a and b to be fully effective (assign weight = 1.0 in potential calculation) only when all angles α_i are equal to, or smaller than 60°. For the cases with $\alpha_i > 60°$, we reduce the weight by weight = $\prod_i (180.0 - \alpha_i)/180.0$ if residue separations between x_i and a, b are ≥2, and at least one of them are ≥7. This procedure can help eliminate the redundant and ineffective interactions in potential derivation and application.

Calculation of ANDIS potential

The ANDIS potential is extracted from an experimental structural dataset of 3519 non-redundant protein chains based on the inverse Boltzmann equation [20]. We assume that the 5 angles (θ_a, θ_b, φ_a, φ_b and χ) are independent of each other at the given distance so as to avoid insufficient statistics. Thus the angle potential can be written as:

$$E^{AG}(\theta_a,\theta_b,\phi_a,\phi_b,\chi\ |r_{a,b}) = -k_{\mathrm{B}}T\ \ln\left[\frac{p^{OBS}(\theta_a,\theta_b,\phi_a,\phi_b,\chi\ |r_{a,b})}{p^{REF}(\theta_a,\theta_b,\phi_a,\phi_b,\chi\ |r_{a,b})}\right] \approx -k_{\mathrm{B}}T\sum\nolimits_i \ln\left\{\frac{p^{OBS}[angle_i(s)\ |r_{a,b}(d)]}{p^{REF}[angle_i(s)\ |r_{a,b}(d)]}\right\} \tag{1}$$

where k_{B} and T are Boltzmann constant and Kelvin temperature, respectively. $r_{a,\ b}$ is the distance between atom type a and b. $angle_i$ is the angle θ_a, θ_b, φ_a, φ_b or χ. $p^{OBS}[angle_i(s)\ |\ r_{a,\ b}(d)]$ and $p^{REF}[angle_i(s)\ |\ r_{a,\ b}(d)]$ are the observed and reference probabilities of $angle_i$ falling into angle bin s at the given distance bin d. The initial count values for each angle bin are set to 0.1. Here we take the average observed value over 12 angle bins as the reference state, which means $p^{REF}[angle_i(s)\ |r_{a,b}(d)] = \sum_{s=1}^{12} p^{OBS}[angle_i(s)\ |\ r_{a,b}(d)]/12$. The observed probabilities are calculated based on the entire structural dataset (3519 non-redundant X-ray structures). Eventually we can obtain an angle-based score matrix with the size of 5 × 167 × 167 × 29 × 12.

Since the best distance cutoff (r_{cut}) is found to be highly depended on the evaluation criteria and the application environments, we make it an adjustable parameter from 7 Å to 15.0 Å for user. Generally, a lower distance cutoff is better for native recognition, while a higher one is favorable for decoy discrimination. The "effective atomic interaction" is employed to enhance the ability of native recognition when $r_{cut} \le 9.0\mathring{A}$. For a distance cutoff of ≥10 Å, the distance-dependent atom-pair potential with random-walk reference state [30] (it yields an additional score matrix of 167 × 167 × 29) is combined with the angle potential to strengthen the ability of decoy discrimination. Therefore, the ANDIS energy score for a given protein sequence S_q with conformation C_p is calculated by

$$E(S_q, C_p) = \begin{cases} \displaystyle\sum_{m=1}^{N\text{-}1}\sum_{n=m+1}^{N} w^{m,n}E^{AG}\left(\theta_a^m,\theta_b^n,\phi_a^m,\phi_b^n,\chi\ |r_{a,b}^{m,n}\right) & if \quad r_{cut} \le 9.5\mathring{A} \\ \displaystyle\sum_{m=1}^{N\text{-}1}\sum_{n=m+1}^{N}\left(0.5\times E^{AG}\left(\theta_a^m,\theta_b^n,\phi_a^m,\phi_b^n,\chi\ |r_{a,b}^{m,n}\right) + E^{RW}\left(r_{a,b}^{m,n}\right)\ \right) & if \quad 10\mathring{A} \le r_{cut} \le 15\mathring{A} \end{cases} \tag{2}$$

where N is the total number of heavy atoms in the protein chain S_q. $r_{a,b}^{m,n}$ is the distance between atom pair m and n (corresponding to atom type a and b, respectively) observed in conformation C_p. r_{cut} is the distance cutoff for $r_{a,b}^{m,n}$, which can be adjusted from 7.0 Å to 15.0 Å by user (Default value: 15.0 Å, and a lower value, e.g. 7.0 Å, is recommended if using ANDIS for native recognition). $w^{m,\ n}$ is the weight for the energy score of atom pair m and n ($w^{m,n} = 1.0\ if\ r_{cut} = 9.5\mathring{A}$), which is determined by the calculation of "effective atomic interactions" (see **Definition of effective atomic interactions**). $E^{RW}(r_{a,b}^{m,n})$ is the distance-dependent atom-pair potential with an ideal random-walk (RW) chain of a rigid step length as the reference state. We calculate RW potential based on the following equation:

$$E^{RW}(r_{a,b}) = -k_{\mathrm{B}}T\ \ln\frac{N^{OBS}(r_{a,b})}{\displaystyle\sum_{p}^{N_{tot}}\left(\frac{r_{a,b}}{r_{cut}}\right)^2\frac{\sum_{n=1}^{L_p}\ \exp\left(-3r_{a,b}^2/2nl^2\right)/n^{3/2}}{\sum_{n=1}^{L_p}\ \exp\left(-3r_{cut}^2/2nl^2\right)/n^{3/2}}N_{a,b}^{OBS,p}(r_{cut})} \tag{3}$$

where $N^{OBS}(r_{a,\ b})$ is the total observed frequencies of atom type pairs (a, b) within a distance bin r to $r + \Delta r$ in the experimental protein dataset. $N_{a,b}^{OBS,p}(r_{cut})$ is the observed frequencies of atom type pairs (a, b) within the distance bin of r_{cut} in protein p. L_p is the sequence length of protein p. l is Kohn length. N_{tot} is the total number of proteins in the experimental dataset. Only atom pairs with residue separation ≥7 are considered. More information about RW potential can be found in the original work by Zhang and Zhang [30].

Decoy datasets for benchmark test

We collected hundreds of decoy sets (each set includes a native structure as well as a bunch of structural decoys) from diverse sources for benchmarking the ANDIS potential (see Table 1). The CASP5–8 decoy sets contain a total of 2759 structures for 143 proteins, which were collected from CASP5-CASP8 experiments by Rykunov and Fiser [43]. The CASP10–13 decoy sets were directly

Table 1 Performance comparison in native recognition

Decoy sets	CASP5–8	CASP10–13	I-TASSER	3DRobot	Rosetta	No. total[d]
No. of targets[a]	143 (2759)	175 (13,474)	56 (24,707)	200 (60,200)	58 (5858)	632 (106,998)
Dfire[b]	64 (0.61)	56 (0.72)	43 (2.80)	1 (0.83)	22 (1.55)	186 (0.99)
RW	65 (1.01)	36 (0.86)	**53** (4.42)	0 (−0.30)	20 (1.48)	174 (0.90)
GOAP	106 (1.67)	89 (1.62)	45 (4.98)	94 (1.85)	45 (3.38)	379 (2.16)
DOOP	135 (1.96)	121 (1.99)	52 (6.18)	197 (3.53)	50 (3.91)	555 (3.02)
ITDA	71 (1.15)	117 (1.67)	52 (4.98)	196 (3.83)	**53** (3.52)	489 (2.70)
VoroMQA	132 (2.00)	111 (1.77)	48 (5.11)	114 (1.89)	43 (3.09)	448 (2.28)
SBROD	88 (1.62)	119 (2.32)	33 (3.25)	49 (1.76)	42 (3.02)	331 (2.13)
AngularQA	59 (1.26)	24 (1.11)	29 (1.82)	9 (0.99)	2 (0.12)	123 (1.08)
ANDIS[c]	**138** (**2.16**)	**129** (**2.32**)	47 (**6.45**)	**200** (**4.99**)	50 (**4.27**)	**564** (**3.67**)

[a] The total number of structures (including native structures) are given in parentheses

[b] The number of proteins whose native structure is given the lowest energy score by the potential are listed outside the parentheses. The average Z-scores of native structures are listed in parentheses. Z-score is defined as $(<E_{decoy}> - E_{\mathrm{native}})/\delta$, where E_{native} is the energy score of native structure, $<E_{decoy}>$ and δ are respectively the average and the standard deviation of energy scores for all decoys in the set. But Z-score for VoroMQA energy score is calculated by $(E_{\mathrm{native}} - <E_{decoy}>)/\delta$, so that Z-scores of native structures for all potentials are "the higher the better"

[c] Calculation is based on a distance cutoff of 7.0 Å

[d] Z-scores are calculated by averaging over all 632 decoy sets

downloaded from http://predictioncenter.org/download_area/. We selected and trimmed these decoy sets based on the following procedure: (i) the prediction sets for targets without experimental structures are removed; (ii) the prediction sets whose target experimental structures are sequentially non-consecutive are removed; (iii) all non-first prediction models (the second to fifth models of predictors) are removed; (iv) the prediction models whose sequences are non-consecutive or shorter than the corresponding experimental structure are removed; (v) all prediction models are trimmed to keep them identical in sequence to the corresponding experimental structure. As a result, the final decoy sets include 175 target proteins (a total of 13,474 structures). The CASP10–13 decoy sets are publicly available at http://qbp.hzau.edu.cn/ANDIS/.

Moreover, we also used other three groups of decoy sets generated by some specific modeling methods. The I-TASSER decoy sets comprise of 56 non-redundant proteins (a total of 24,707 structures) whose structure decoys were generated by I-TASSER Monte Carlo simulations [44] and refined by GROMACS4.0 MD simulation [45]. The 3DRobot decoy sets were generated by a specialized decoy generating method we previously developed [35], which include 200 non-redundant proteins (a total of 60,200 structures). The Rosetta decoy sets include a total of 5858 structures for 58 proteins, which were generated by Rosetta ab initio structure prediction [46].

Other potentials for benchmark comparison

We benchmarked ANDIS with other 8 state-of-the-art potentials. Two of them (Dfire [36] and RW [30]) are purely distance-dependent atom-pair statistical potentials with different analytical assumptions of reference state. GOAP [37] depends on the relative orientation of the planes associated with each heavy atom in interacting pairs, which combines Dfire with an angle-dependent potential. ITDA [47] integrates the distance-dependent atom-pair potential with a new component for estimating the backbone conformational entropies. VoroMQA [38] combines the idea of statistical potentials with the use of interatomic contact areas instead of distances. Contact areas, derived using Voronoi tessellation of protein structure, are capable of capturing both explicit interactions between protein atoms and implicit interactions of protein atoms with solvent. The other 3 potentials (DOOP [40], SBROD [48] and AngularQA [49]) employ machine learning methods to different extent. DOOP is a neural network-based potential with distance distributions of different atom pairs as input features. It also includes a torsion potential term which describes the local conformational preference. SBROD is trained based on Ridge Regression with four different structural features: residue-residue orientations, contacts between backbone atoms, hydrogen bonding, and solvent-solute interactions. AngularQA is derived based on Long Short-Term Memory (LSTM) network with the angles between residues being the core features. Like ANDIS, all the 8 potentials are single-model quality assessment methods.

Results

Effects of distance cutoff on ANDIS's performance

Distance cutoff is one of the most essential parameter for distance-dependent potentials. A series of distance cutoffs (from 5.8 Å to 16.0 Å) were tested to derive different versions of ANDIS potential. Figure 2 shows their average performance over all 632 decoy sets. Potential based on distance cutoff of around 7.0 Å achieves the

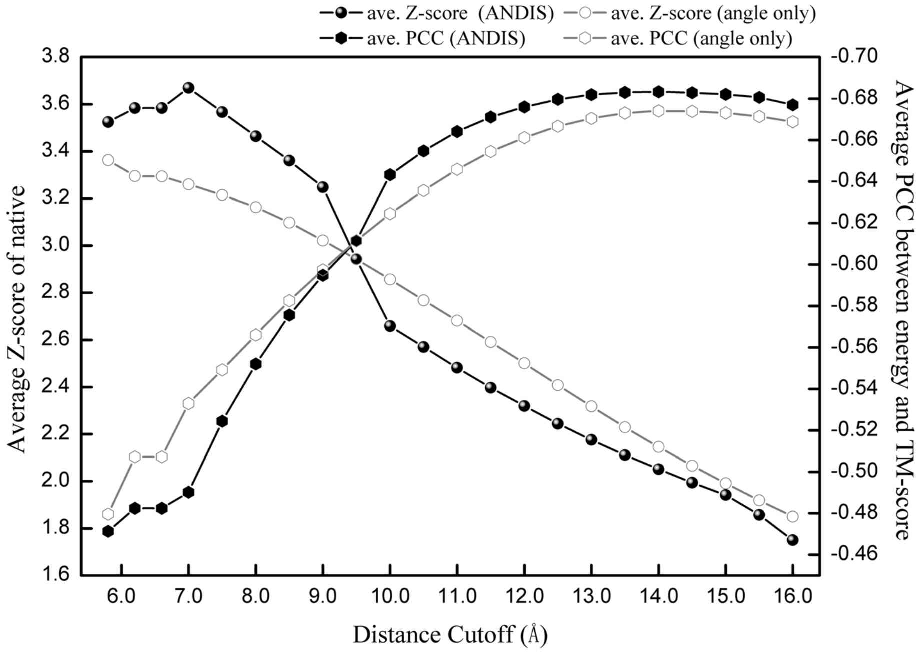

Fig. 2 Effects of distance cutoff on ANDIS's performance. The results are averaged over all 632 structural decoy sets. "angle only" refers to the pure angle potential without involvement of "effective atomic interaction" and distance-dependent atom-pair potential. Since lower energy score (higher TM-score) is desired, the value of PCC is negative, the lower the better

highest average Z-score (of native structure). Afterwards, the average Z-score decreases linearly with the increase of distance cutoff. However, the average PCC (between ANDIS energy and TM-score) varies with distance cutoff in the opposite trend. These results indicate that the potential's abilities of native recognition and decoy discrimination cannot be optimized simultaneously with the same distance cutoff. Generally, a lower distance cutoff is better for native recognition, while a higher one is favorable for decoy discrimination. But the optimal distance cutoff for decoy sets from different sources may vary. As shown in Additional file 1: Figure S1, the best cutoff of native recognition for I-TASSER decoy sets is 9.0 Å, and the best cutoff of decoy discrimination for 3DRobot decoy sets is 10.0 Å. Therefore, ANDIS provides distance cutoff as an adjustable parameter from 7.0 Å to 15.0 Å with bin-width of 0.5 Å. The default value is set to 15.0 Å in favor of decoy discrimination, and 7.0 Å is recommended for native recognition.

Since the "effective atomic interaction" is beneficial for native recognition but unhelpful for decoy discrimination, we include it only when a lower distance cutoff (≤ 9.0 Å) is adopted. As shown in Fig. 2, the average Z-score is significantly improved compared with that of angle potential only. The results for cases with higher distance cutoff (≥ 10.0 Å) also demonstrate a remarkable promotion in decoy discrimination achieved by incorporation of the distance-dependent atom-pair potential with random-walk reference state.

Moreover, we also checked the distance cutoffs used by the distance-dependent potentials listed in Table 1 (Dfire, RW, GOAP and DOOP), and found that most of them are around 15 Å, except that of DOOP (6.5 Å). This could provide a possible explanation for DOOP's outstanding performance in native recognition.

Performance comparison in native recognition

We applied ANDIS as well as other 8 potentials on the 632 decoy sets from CASP experiments [50], I-TASSER [30], 3DRobot [35] and Rosetta [46]. Table 1 summarizes the performances of different potentials in native recognition (recognize the native structure among a set of structural decoys). ANDIS (distance cutoff of 7.0 Å is used) recognizes 564 native structures (success rate is about 90%) and achieves an average Z-score of 3.67 over all decoy sets, which is remarkably better than that of the other eight potentials. For the CASP5–8 [43], CASP10–13 and 3DRobot decoy sets, ANDIS has the best performances. For I-TASSER and Rosetta decoy sets, ANDIS fails to achieve the best success rate, but still has the best Z-score.

The atomic distance-dependent pair-wise potentials, Dfire and RW, perform much worse than other potentials. Although their capabilities for native recognition can be remarkably improved by adjusting the distance cutoff and residue interval [39], they failed to outperform DOOP and

ANDIS (data not shown). GOAP significantly outperforms Dfire and RW, but still has large gaps compared with other 4 potentials. The neural network-based potential DOOP (with distance cutoff of 6.5 Å) is the only one with comparable performance to ANDIS. Moreover, ITDA and VoroMQA, the two recently developed statistical potentials, both underperform DOOP in native recognition. However, ITDA achieves the best success rate (53 out of 58) on Rosetta decoy sets. The other two machine learning-based methods, SBROD and AngularQA, perform much worse than DOOP in native recognition, which is possibly because they are mainly designed for decoy ranking.

Performance comparison in decoy discrimination

The more practical use of statistical potential is to discriminate between good and bad structural decoys. Table 2 summarizes the performances of different potentials in decoy discrimination. We evaluate the ability of decoy discrimination based on the average Pearson's correlation coefficient (PCC) between energy score and TM-score, as well as the 20% enrichment which measures the relative occurrence of the most accurate (by TM-score) 20% decoys among the 20% best scoring (by potential) decoys. The outstanding performances of SBROD on CASP decoy sets help it achieves the best average performances over all decoy sets. However, its performances on the rest three groups of decoy sets are far worse than those of other methods (except AngularQA). In fact, SBROD are trained directly based on CASP5-CASP10 datasets, which probably brings it an inherent bias to CASP decoy sets. ANDIS achieves both the best average PCC (– 0.681) and the best average 20% enrichment (2.83) over all 632 decoy sets (except SBROD). The performances of VoroMQA are relatively close to that of ANDIS. GOAP outperforms all other potentials on 3DRobot decoy sets. In fact ANDIS is able to surpass GOAP on 3DRobot decoy sets if a distance cutoff between 10.0 Å to 13.0 Å is adopted (e.g., the average PCC and 20% enrichment on 3DRobot decoy sets are 0.910 and 4.14 when distance cutoff is set to 10.0 Å). DOOP and ITDA, which are outstanding in native recognition, perform noticeably worse than other potentials in decoy discrimination (except AngularQA). The bad performances of AngularQA are probably because it is mainly designed to serve as an energy component, not a standalone QA method.

Calculation by GDT_TS (instead of TM-score) came up with very similar results (data not shown).

Discussion

Effects of protein dataset on ANDIS's performance

By the beginning of 2018, the total number of structures deposited in the Protein Data Bank [51] has almost reached 140,000. The size and scope of protein dataset are no longer a problem for potential derivation. To demonstrate the correlation between dataset size and ANDIS's performance, we derived ANDIS based on different number of protein structures from the dataset (3519 X-ray structures). As shown in Fig. 3, the average Z-score of native increases with the size of protein dataset, faster when the dataset is relatively small (e.g., < 1200), stabilized gradually when the dataset size exceeds 2000. However, the average PCC is very insensitive to the size of dataset. It is noteworthy that the potential based on only 400 structures can already achieve an average PCC very close to the optimal. This implies that the rest 3000 structures actually have very little contribution to promote potential's ability of decoy discrimination. The same procedure was also conducted on other

Table 2 Performance comparison in decoy discrimination

Decoy sets	CASP5–8	CASP10–13	I-TASSER	3DRobot	Rosetta	Average[e]
Dfire[b]	–0.548 (2.16)	–0.441 (2.01)	– 0.480 (**1.62**)	–0.860 (3.77)	– 0.366 (1.97)	–0.594 (2.56)
RW	–0.550 (2.16)	–0.462 (2.01)	– 0.476 (1.59)	–0.863 (3.80)	– 0.361 (1.95)	–0.601 (2.57)
GOAP	–0.607 (2.66)	–0.550 (2.13)	– 0.473 (1.61)	**–0.900** (**4.04**)	**– 0.406** (1.99)	–0.654 (2.79)
DOOP	–0.442 (1.95)	–0.415 (1.90)	– 0.333 (1.41)	–0.874 (4.00)	– 0.285 (1.67)	–0.547 (2.51)
ITDA	–0.392 (2.03)	–0.452 (2.03)	– 0.431 (1.54)	–0.841 (3.70)	– 0.302 (1.68)	–0.545 (2.48)
VoroMQA[c]	0.665 (2.66)	0.628 (2.26)	0.450 (1.44)	0.893 (3.91)	0.366 (1.86)	0.680 (2.76)
SBROD[c]	**0.793** (**3.06**)	**0.831** (**2.26**)	0.397 (1.49)	0.857 (3.34)	0.270 (1.62)	**0.741** (2.66)
AngularQA[c]	0.441 (0.122)	0.426 (0.304)	0.323 (0.579)	0.543 (0.224)	0.042 (0.961)	0.422 (0.32)
ANDIS[d]	–0.663 (2.80)	–0.607 (2.18)	**–0.503** (1.59)	– 0.891 (3.95)	–0.401 (**2.05**)	– 0.681 (**2.83**)

[a] the native structures in the decoy sets are ignored when calculating PCC and "20% enrichment"

[b] The average Pearson's correlation coefficient between energy and TM-score (PCC) is listed outside the parentheses. The average value of 20% enrichment is listed in parentheses. "20% enrichment" means the relative occurrence of the most accurate (by TM-score) 20% models among the 20% best scoring (by potential) models compared to that for the entire decoy set. The possible value of 20% enrichment ranges from 0 to 5, the higher the better

[c] Since the energy scores of VoroMQA, SBROD and AngularQA are the higher the better, the PCC between them and TM-score is positive

[d] Calculation is based on a distance cutoff of 15.0 Å

[e] by averaging over all 632 decoy sets

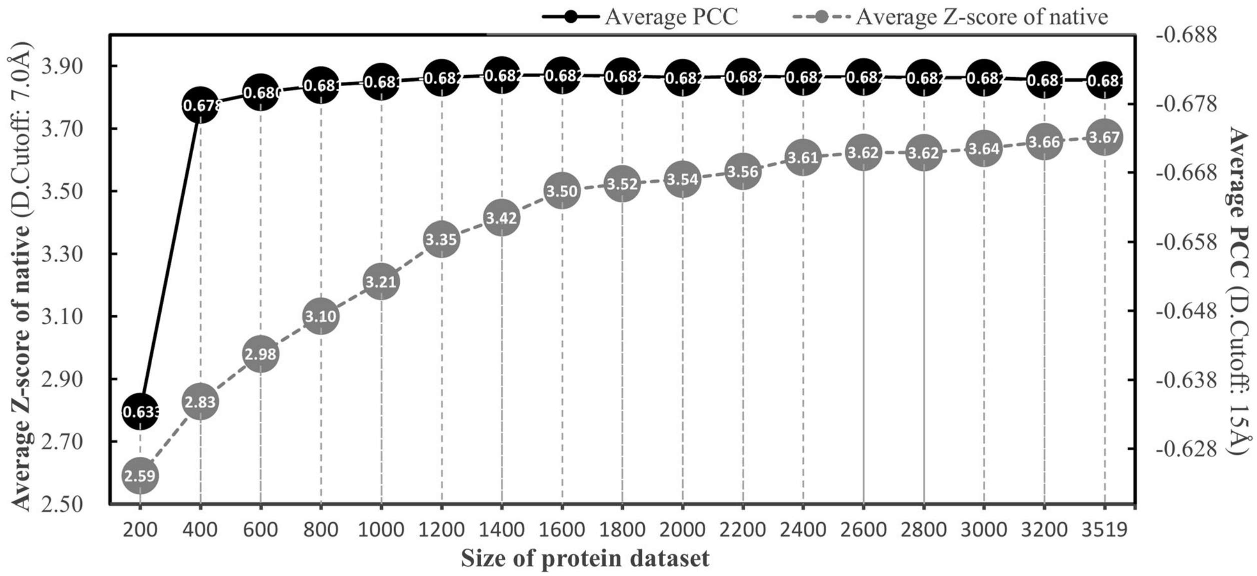

Fig. 3 Overall effects of dataset size on ANDIS's performance. ANDIS is re-extracted based on different number of structures from the original dataset (3519 structures)

datasets listed in Additional file 1: Figure S2, similar trends were observed. In general, a dataset with around 3000 structures is adequate for ANDIS to obtain the optimal or near-optimal performance in native recognition.

Moreover, on what basis should a protein dataset be determined, and how does the choice of dataset affect potential's performances? Here we prepared a series of structure datasets according to the pre-compiled PDB lists for various parameter sets (resolution, sequence identity, etc.) from PISCES [41]. We derived the ANDIS potential based on different datasets and summarized the test results in Additional file 1: Figure S2. It is easy to see that the performance variation brought by dataset with different parameter sets is very limited. There are almost no changes on average PCC for all 5 groups of decoy sets. The average Z-score for 3DRobot decoy sets increases slightly with the decrease of dataset size, but reverse trends can be seen for I-TASSER and Rosetta decoy sets. In fact, results based on datasets with size > 3000 are relatively stable.

What kind of native structures are hard to be recognized?

Although 90% of native structures are successfully recognized by ANDIS, what are the other unrecognized 10%? We checked all the 58 unrecognized native structures, and found that their average length is significantly lower than that of the recognized. We also calculated the MolProbity score [52] of native structure. It is a well-known metric for estimating the physical reasonableness of protein structure. Figure 4 shows the length and MolProbity score of all 175 native structures in CASP10–13 decoy sets. We can see that all 9 native structures with length < 65 residues and 75% (24 out of 32) of native structures with MolProbity score > 2.0 are not recognized by ANDIS. Quite the contrary, more than 90% of native structures with length > 65 and MolProbity score < 2.0 are successfully recognized by ANDIS. Since higher MolProbity score implies worse structural quality (or lower resolution), these observations indicate that the hard targets for native recognition have a certain degree of commonality. In another sense, for the target protein of small size (or target protein whose experimental structure has relatively low resolution), current prediction methods are capable of generating protein models comparable to the experimental structure. Furthermore, all native structures in I-TASSER and Rosetta decoy sets are small proteins with average lengths of 80 residues and 83 residues, respectively. There is no evident difference in length between the recognized and the unrecognized native structures from them. But the average MolProbity scores of the unrecognized native structures from I-TASSER and Rosetta decoy sets are 2.386 and 2.506 respectively, much larger than those of the recognized native structures from them (1.223 and 1.771, respectively). Similar results are observed in CASP5–8 decoy sets. In fact all the 5 unrecognized native structures from CASP5–8 decoy sets are ranked second by ANDIS, only inferior to one prediction model.

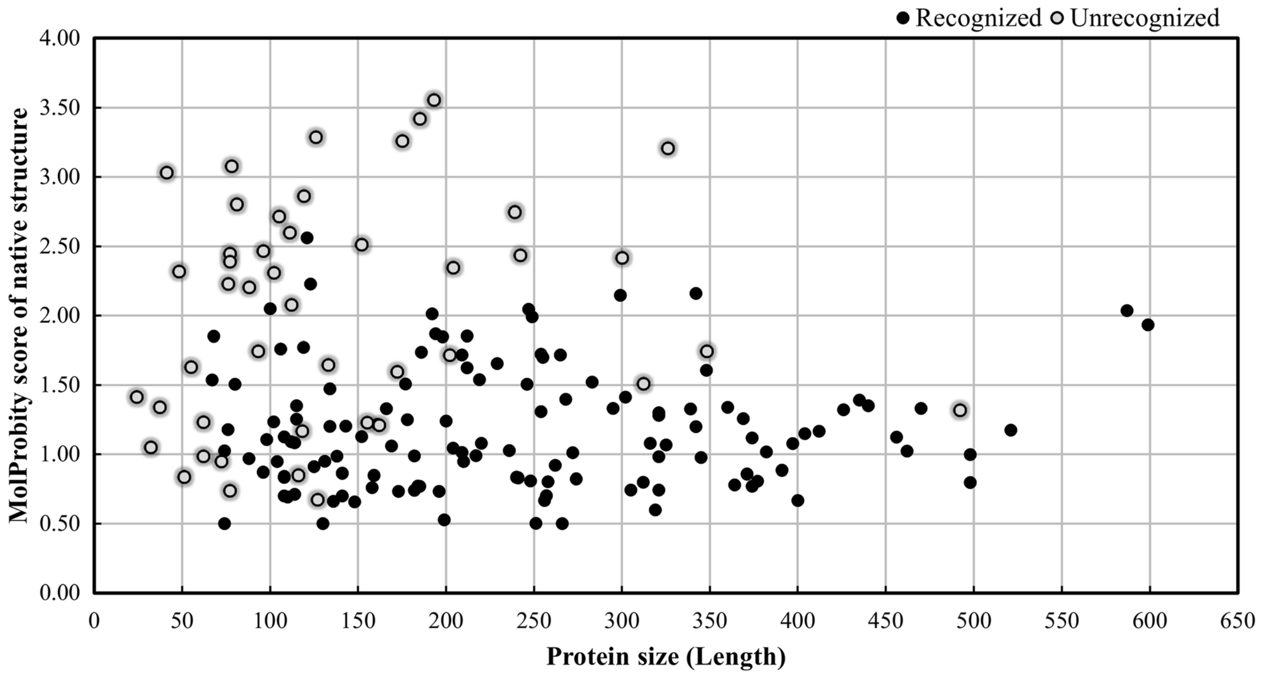

Fig. 4 The protein size and MolProbity score for native structures in CASP10-13 decoy sets. ANDIS recognized 129 (out of 175) native structures in CASP10-13 decoy sets. The 46 unrecognized native structures are highlighted by shade open circles

Conclusions

Our study demonstrates that distance cutoff plays a crucial role in distance-dependent statistical potential. Generally, a lower distance cutoff is better for native recognition, while a higher one is favorable for decoy discrimination. We developed an atomic angle- and distance-dependent potential (ANDIS) with distance cutoff being an adjustable parameter. ANDIS's ability of native recognition is remarkably promoted by introducing the "effective atomic interactions". Most of the native structures that fail to be recognized are small proteins or with poor MolProbity score. A distance-dependent atom-pair potential with random-walk reference state is combined to ANDIS when distance cutoff is ≥10 Å, which successfully enhances ANDIS's ability of decoy discrimination. The results of benchmark tests indicate that ANDIS outperforms other state-of-the-art potentials in both native recognition and decoy discrimination.

Moreover, we investigated the effects of protein dataset on potential's performance. Datasets culled by different parameter sets don't make a real difference on ANDIS's performance, but the size of dataset should reach a certain level. A dataset with about 3000 structures is adequate for ANDIS to achieve the optimal performance in native recognition. While the size reduces to hundreds of structures for optimizing the ability of decoy discrimination. Why is there such a difference? What is the best size of a representative dataset? How is the limitation of a potential in information extraction? These interesting questions remain to be further explored.

Additional file

Additional file 1: Figure S1. Effects of distance cutoff on ANDIS's performance for different decoy sets. **Figure S2.** Effects of protein dataset on ANDIS's performance for different decoy sets. (DOCX 222 kb)

Abbreviations

CASP: The Critical Assessment of protein Structure Prediction experiments; PCC: the Pearson's correlation coefficient

Acknowledgements

Not applicable.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 11604111, No. 11675060 and No. 91730301), the Huazhong Agricultural University Scientific and Technological Self-innovation Foundation Program (Grant No.2015RC021), and the fundamental Research Funds for the Central-Universities (Grant No.2662018JC017). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The standalone package of ANDIS, the non-redundant structure dataset and the CASP10–13 decoy sets are publicly available at http://qbp.hzau.edu.cn/ANDIS/.

Authors' contributions
HD conceived and designed the study and wrote the manuscript. ZY and MY carried out the calculations. YY prepared the structural decoy data. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Physics, College of Science, Huazhong Agricultural University, Wuhan 430070, China. [2]Institute of Applied Physics, Huazhong Agricultural University, Wuhan 430070, China.

Received: 28 November 2018 Accepted: 13 May 2019
Published online: 03 June 2019

References

1. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins-structure Function. Bioinformatics. 1995;21(3):167–95.
2. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J. 2003;85(2):1145–64.
3. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545–614.
4. 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(16):1668–88.
5. Bhattacharya D, Cao R, Cheng J. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics. 2016;32(18):2791–9.
6. Misura KMS, David B. Progress and challenges in high-resolution refinement of protein structure models. Proteins: Struct, Funct, Bioinf. 2005;59(1):15–29.
7. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure. 2011;19(12):1784–95.
8. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J. 2011;101(10):2525–34.
9. Bhattacharya D, Nowotny J, Cao R, Cheng J. 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic Acids Res. 2016; 44(W1):W406–9.
10. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008;71(1):261–77.
11. Roche DB, Buenavista MT, McGuffin LJ. Assessing the quality of modelled 3D protein structures using the ModFOLD server. Methods Mol Biol. 2014; 1137:83–103.
12. Uziela K, Menendez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model quality assessments using deep learning. Bioinformatics. 2017;33(10):1578–80.
13. Mackerell AD Jr. Empirical force fields for biological macromolecules: overview and issues. J Comput Chem. 2004;25(13):1584–604.
14. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008;18(3):342–8.
15. Senn HM, Thiel W. QM/MM methods for biomolecular systems. Angew Chem Int Ed Eng. 2009;48(7):1198–229.
16. 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(1): 288–301.
17. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006;15(11):2507–24.
18. Deng H, Jia Y, Wei Y, Zhang Y. What is the best reference state for designing statistical atomic potentials in protein structure prediction? Proteins. 2012;80(9):2311–22.
19. Cao R, Bhattacharya D, Hou J, Cheng J. DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics. 2016;17(1):495.
20. 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(4):859–83.
21. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995;5(2):229–35.
22. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998;275(5):895–916.
23. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci U S A. 2003; 100(6):3215–20.
24. Faraggi E, Xue B, Zhou YQ. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins. 2009;74(4): 847–56.
25. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci U S A. 2000;97(6):2550–5.
26. Berrera M, Molinari H, Fogolari F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics. 2003;4(1):1–26.
27. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001;44(3): 223–32.
28. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins. 2015;41(1):40–6.
29. Zhao F, Xu J. A position-specific distance-dependent statistical potential for protein structure and functional study. Structure. 2012;20(6):1118–26.
30. 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(10):e15386.
31. Liang S, Zhou Y, Grishin N, Standley DM. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. J Comput Chem. 2011;32(8):1680–6.
32. Samudrala R, Levitt M. Decoys 'R'Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci. 2000;9(07):1399–401.
33. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. 2003; 31(14):3982–92.
34. Topf M, Baker ML, John B, Chiu W, Sali A. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. J Struct Biol. 2005;149(2):191–203.
35. Deng H, Jia Y, Zhang Y. 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics. 2016;32(3):378–87.
36. 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(11):2714–26.
37. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J. 2011;101(8): 2043–52.
38. Olechnovic K, Venclovas C. VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins. 2017;85(6):1131–45.
39. Yao Y, Gui R, Liu Q, Yi M, Deng H. Diverse effects of distance cutoff and residue interval on the performance of distance-dependent atom-pair potential in protein structure prediction. BMC Bioinformatics. 2017;18(1):542.
40. Chae MH, Krull F, Knapp EW. Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. Proteins. 2015;83(5):881–90.
41. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589–91.
42. Ferrada E, Melo F. Effective knowledge-based potentials. Protein Sci. 2009; 18(7):1469–85.
43. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics. 2010; 11(1):128.
44. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725–38.

45. Hess B, Kutzner C, Van Der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput. 2008;4(3):435–47.
46. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins. 2003;53(1):76–87.
47. Wang X, Zhang D, Huang SY. New Knowledge-Based Scoring Function with Inclusion of Backbone Conformational Entropies from Protein Structures. J Chem Inf Model. 2018;58(3):724–32.
48. Karasikov M, Pagès G, Grudinin S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics. Oxford: University Press (OUP). pp.1–8. https://doi.org/10.1093/bioinformatics/bty1037.
49. Conover M, Staples M, Si D, Sun M, Cao R. AngularQA: protein model quality assessment with LSTM networks; 2019.
50. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005;15(3):285–9.
51. 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(1):235–42.
52. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB 3rd, Snoeyink J, Richardson JS, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007;35(Web Server):W375–83.