Structure
Article

A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study

Feng Zhao[1] and Jinbo Xu[1,*]
[1]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA
*Correspondence: jinboxu@gmail.com
DOI 10.1016/j.str.2012.04.003

SUMMARY

Although studied extensively, designing highly accurate protein energy potential is still challenging. A lot of knowledge-based statistical potentials are derived from the inverse of the Boltzmann law and consist of two major components: observed atomic interacting probability and reference state. These potentials mainly distinguish themselves in the reference state and use a similar simple counting method to estimate the observed probability, which is usually assumed to correlate with only atom types. This article takes a rather different view on the observed probability and parameterizes it by the protein sequence profile context of the atoms and the radius of the gyration, in addition to atom types. Experiments confirm that our position-specific statistical potential outperforms currently the popular ones in several decoy discrimination tests. Our results imply that, in addition to reference state, the observed probability also makes energy potentials different and evolutionary information greatly boost performance of energy potentials.

INTRODUCTION

Millions of protein sequences are publicly available, but a large percentage of them lack of solved structures, which are essential to the understanding of their molecular mechanisms and to many important applications. Elucidating the structure and function of a protein often requires an accurate physics-based or knowledge-based potential to quantify interactions among residues or atoms. Previous studies (Bradley et al., 2005; Skolnick, 2006) indicate that knowledge-based statistical potentials (Lu et al., 2008; Miyazawa and Jernigan, 1985; Shen and Sali, 2006; Simons et al., 1997; Sippl, 1990; Tanaka and Scheraga, 1976; Zhang and Zhang, 2010; Zhou and Zhou, 2002) compare favorably to physics-based potentials (Brooks et al., 2009; Bryngelson et al., 1995; Case et al., 2005; Dill, 1985, 1997; Dobson et al., 1998; Schuler et al., 2001; Shakhnovich, 2006) in many applications including ab initio folding (Jones and Thirup, 1986; Kihara et al., 2001; Levitt, 1992; Simons et al., 1997; Wu et al., 2007a; Zhao et al., 2008, 2010), docking (Zhang et al., 1997), binding (Kortemme and Baker, 2002; Laurie and Jackson, 2005), mutation study (Gilis and Rooman, 1996, 1997), decoy ranking (Bauer and Beyer, 1994; Casari and Sippl, 1992; Gatchell et al., 2000; Hendlich et al., 1990; Samudrala and Moult, 1998; Simons et al., 1999; Vendruscolo et al., 2000), and protein model quality assessment (Jones and Thornton, 1996; Panchenko et al., 2000; Peng and Xu, 2010; Reva et al., 1997; Sippl, 1993). Knowledge-based statistical potentials extract interactions directly from the solved protein structures in the Protein Data Bank (PDB) (Kouranov et al., 2006) and are simpler and easier to use than physics-based potentials. A lot of statistical potentials have been developed including the widely used DOPE (Shen and Sali, 2006) and DFIRE (Zhou and Zhou, 2002). Some statistical potentials quantify local atomic interactions(e.g., torsion angle potential), while others capture nonlocal atomic interactions (e.g., distance-dependent potential).

A lot of statistical potentials are derived from the inverse of the Boltzman law. In the traditional position-independent, distance-dependent statistical potentials (e.g., DOPE and DFIRE), the interaction potential of two atom types a and b can be estimated as follows:

$$U(d|a,b) = -kT\,ln\frac{P(d|a,b)}{q(d)}, \quad (1)$$

where k is the Boltzman constant, T is the temperature, and d represents the interatom distance shell $[d, d+\Delta d]$. Meanwhile, $P(d|a,b)$ is the observed probability of two atoms interacting within the distance shell, and $q(d)$ is the reference state (i.e., the expected probability of two noninteracting atoms within the distance shell). The reference state is used to rule out the average and generic correlation of two atoms not due to atomic interactions. Most statistical potentials parameterize the observed atomic interacting probability by (residue-specific) atom types and use a simple counting method to estimate it. For example, $P(d|a,b)$ in Equation (1) is often calculated by $count(d,a,b)/\sum_d count(d,a,b)$, where $count(d,a,b)$ is the number of observed occurrences of two atoms a and b within a distance shell $[d, d+\Delta d]$. The distance-dependent statistical potentials developed so far mainly differ from one another in estimating the reference state (Shen and Sali, 2006; Wu et al., 2007b; Zhang and Zhang, 2010; Zhou and Zhou, 2002). Some (e.g., DFIRE and DOPE) use analytical methods to estimate the reference state, while others use statistical methods such as KBP (Lu and Skolnick, 2001) and RAPDF (Samudrala and Moult, 1998). Although using different reference states, these potentials do not have very different energy curves (see Figure 2 in Zhang and Zhang, 2010; and Figure 4 in Shen and Sali, 2006). These

traditional position-independent potentials share a couple of common properties: (1) Once the atom distance and types are given, the atomic interaction potential is fixed across all proteins and residues; and (2) the atomic interaction potentials approach to 0 when the distance is larger than 8 Å.

This article presents a protein-specific and position-specific statistical potential evolutionary pairwise distance-dependent potential (EPAD). We parameterize the observed probability in EPAD by the evolutionary information and radius of gyration of the protein under consideration, in addition to atom types. EPAD distinguishes itself from others in that it may have different energy profiles for two atoms of given types, depending on the protein under consideration and the sequence profile context of the atoms (i.e., evolutionary information). Evolutionary information has been extensively used in protein secondary structure prediction(Jones, 1999; Wang et al., 2011), fold recognition (Maiorov and Crippen, 1992; Panchenko et al., 2000; Peng and Xu, 2009, 2010; Sippl and Weitckus, 1992; Skolnick et al., 2000), protein alignment (Notredame et al., 2000; Pei et al., 2008; Wu and Zhang, 2008b; Xu, 2005; Zhang and Skolnick, 2005), model quality assessment (Jones and Thornton, 1996; Panchenko et al., 2000; Peng and Xu, 2010; Reva et al., 1997; Sippl, 1993), and even protein conformation sampling (Bystroff et al., 2000; Simons et al., 1997; Zhao et al., 2008, 2010). However, evolutionary information is rarely used to design a statistical potential suitable for ab initio protein folding. Panjkovich et al. (2008) have developed a structure-specific statistical potential using evolutionary information for the assessment of comparative models. Nevertheless, this potential is not position specific and subject to a couple of restrictions: (1) It requires the knowledge of at least one native structure in a protein family, so it cannot be applied to ab initio folding a protein with novel fold or to the assessment of models built from distantly related templates; and (2) it requires at least 50 sequence homologs for sufficient statistics. By contrast, our statistical potential is not subject to such restrictions and, thus, is more widely applicable. We term our statistical potential EPAD.

Experimental results show that our position-specific statistical potential outperforms many currently popular ones in several decoy discrimination tests. These results imply that, in addition to reference state, the observed atomic interacting probability is also critical to statistical potentials and can be estimated much more accurately using context-specific evolutionary information.

RESULTS AND DISCUSSION

Overview

Let a_i and a_j denote two atoms of two residues at positions i and j, respectively. Let S_i and S_j denote position-specific sequence profile contexts at positions i and j, respectively (see Experimental Procedures for the definition of sequence profile contexts). Our distance-dependent statistical potential is defined as follows:

$$U(d|a_i,a_j,S_i,S_j,r_g) = -kT\,ln\frac{P(d|a_i,a_j,S_i,S_j,r_g)}{q(d|r_g)}, \qquad (2)$$

where k is the Boltzman constant, T is the temperature, $q(d|r_g)$ is the reference state, and $P(d|a_i,a_j,S_i,S_j,r_g)$ is the observed probability of two atoms a_i and a_j interacting within a distance shell $[d,d+\Delta d]$ conditioned on atom types, residue sequence profile contexts, and r_g (the estimated radius of gyration of the protein under consideration). We use $r_g = 2.2N^{0.38}$ to estimate the radius of gyration where N is the protein sequence length. In comparison to Equation (1), our statistical potential differs from the traditional position-independent potentials (e.g., DOPE and DFIRE) in a couple of aspects. First, the interaction potential of two atoms is protein specific since it depends on the evolutionary information and radius of gyration of the protein under consideration. Second, our potential is position specific since it is parameterized by sequence profile contexts in addition to atom types. We use the same reference state as DOPE (Shen and Sali, 2006), which is a finite sphere of uniform density with appropriate radius. That is, the reference state depends on only the size of a sample protein structure (see Supplemental Experimental Procedures, available online, for more details).

We cannot use the simple counting method to calculate $P(d|a_i,a_j,S_i,S_j,r_g)$ since there is an insufficient number of solved protein structures in PDB for reliable simple counting of sequence profile contexts S_i and S_j. Instead, we apply a probabilistic neural network (PNN) (Specht, 1990) to estimating $P(d|a_i,a_j,S_i,S_j,r_g)$ when both a_i and a_j are C_α atoms. PNN will effectively learn the sophisticated relationship between interatom distance and sequence profiles and yield accurate distance probability distribution. We then estimate $P(d|a_i,a_j,S_i,S_j,r_g)$ for non-C_α atoms conditioned upon C_α distance distribution.

Distance Dependence of the Statistical Potentials

To examine the difference between our potential EPAD and the popular DOPE, we plotted the potentials as a function of interatom distance for two atom pairs, as shown in Figure 1. Figure 1A shows the DOPE interaction potential for the atom pair ALA C_α and LEU C_α and also the EPAD interaction potential for this pair in three different positions of protein 1gvp. DOPE has the same energy curve for this atom pair regardless of its sequence positions. In particular, DOPE always has a favorable potential when the distance of this pair of atoms is between 5 and 7 Å and has an interaction potential close to 0 when the distance is larger than 8.0 Å. By contrast, EPAD has one unique energy curve for this atom pair for each position. The figure legend indicates the corresponding native distances between atom ALA C_α and atom LEU C_α at the three different sequence positions. For example, the bottom curve in Figure 1A visualizes the EPAD interaction potential for the ALA C_α and LEU C_α pair with native distance 8.31 Å. This curve shows that when the distance between ALA C_α and LEU C_α is close to the native, their EPAD interaction potential is favorable. In fact, EPAD always has a favorable potential for these three ALA C_α and LEU C_α pairs when their distances are close to the natives.

Figure 1B compares the EPAD and DOPE interaction potentials for another atom pair Cys N and Trp O in three different proteins of 1B3A, 1BKR, and 1PTQ. Similar to Figure 1A, EPAD has different interaction potentials for the same atom pair in three different proteins, while DOPE has the same potential across all proteins. In particular, EPAD has a favorable potential when the distance between Cys N and Trp O is close to the native. Nevertheless, DOPE has a favorable potential when their

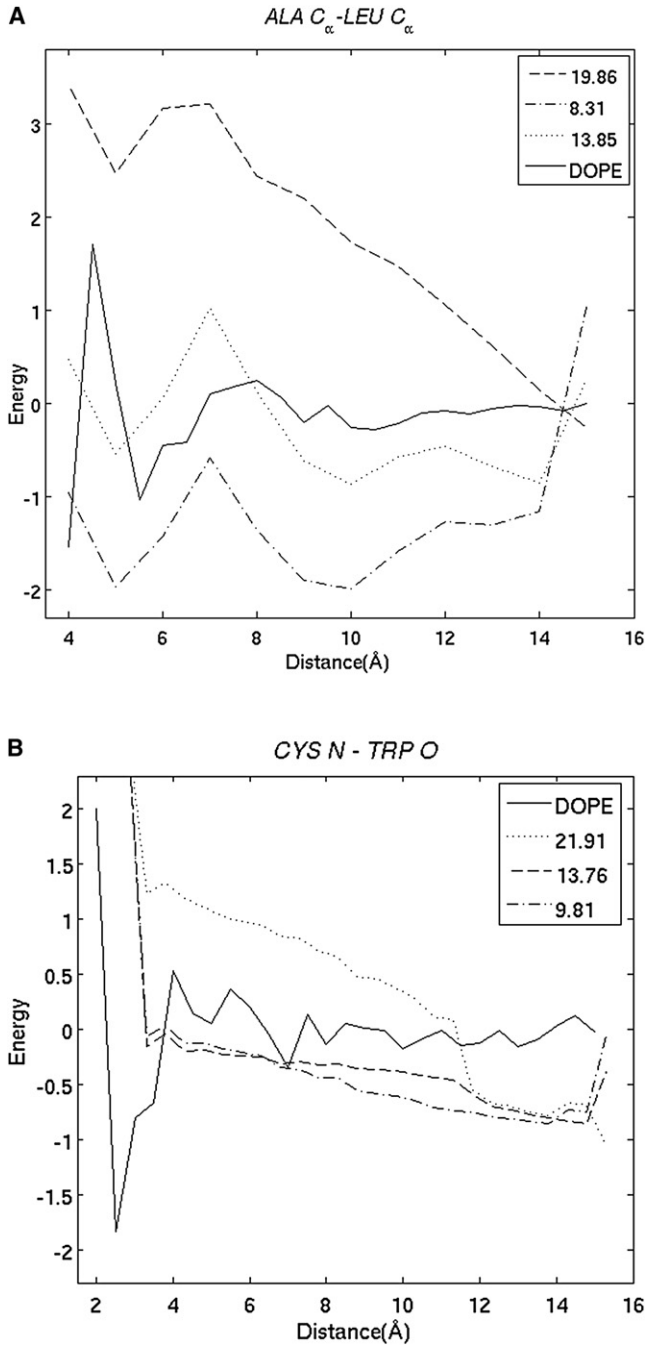

Figure 1. Distance Dependence of DOPE and Our Potential EPAD

(A) The solid curve shows the DOPE interaction potential for atom C_α in ALA and atom C_α in LEU. The other three curves show the EPAD potentials for the same atom pair in three different positions of protein 1gvp. The legend shows the native distances of this atom pair in these positions.

(B) The curves show the DOPE and EPAD potentials for atom N in Cys and atom O in Trp in three different proteins of 1B3A, 1BKR, and 1PTQ.

distance is between 2 and 4 Å and a potential close to 0 when the distance is >8.0 Å.

In summary, our statistical potential EPAD is significantly different from currently popular potentials such as DOPE and DFIRE. DOPE, DFIRE, RAPDF, and RW have more or less similar energy profiles for atom pairs of the same type. The difference between EPAD and DOPE, DFIRE, RAPDF, or RW is much larger than those among DOPE (Shen and Sali, 2006), DFIRE (Zhou and Zhou, 2002), RAPDF (Samudrala and Moult, 1998), and RW (Zhang and Zhang, 2010).

Table 1. Performance of EPAD and Several Popular Full-Atom Statistical Potentials on the Rosetta Decoy Set

	EPAD	DOPE	DFIRE	OPUS	RW	EPAD2
No. of natives identified	34	21	21	39	21	**46**
Ranking of native	15.7	23.7	21.6	9.8	23.9	**13.4**
First-ranked GDT score	51.6	49.7	49.4	49.7	48.5	**52.4**
Pearson CC	**−0.42**	−0.32	−0.25	−0.20	−0.32	−0.39
Z score	−2.46	−1.61	−1.67	**−3.27**	−1.51	**−3.28**

Numbers in bold indicate the best performance. The per-target Pearson CC is calculated between the energy value and GDT score, and then the average value is reported in the table.

Performance on Decoy Discrimination

We tested our backbone-based potential EPAD on several decoy sets including the Rosetta set (Qian et al., 2007), the CASP9 models, the I-TASSER data set (Zhang and Zhang, 2010), the CASP5-8 data set (Rykunov and Fiser, 2010), and the Decoy 'R' Us (Samudrala and Levitt, 2000) set, as well as an in-house large set of template-based models. We evaluated EPAD and several others (DOPE, DFIRE, OPUS [Lu et al., 2008] and RW [Zhang and Zhang, 2010]) using five performance metrics including the number of correctly identified natives, the ranking of the native structures, the Z score of the native energy, the model quality of the first-ranked decoys, and the Pearson correlation coefficient (Pearson CC) between the energy and the model quality. The first three metrics evaluate how well a statistical potential can differentiate natives from decoys. The Pearson CC is more important when we want to apply the potentials to folding a protein sequence. We evaluated the model quality of a decoy using the widely used Global Distance Test (GDT; Zemla, 2003; Zemla et al., 1999, 2001), which compares a decoy with its native and generates a quality value between 0 and 100. The higher the GDT score, the better quality the decoy has.

Performance on the 2007 Rosetta Data Set

The set contains decoys generated and refined by the popular fragment assembly ab initio folding program Rosetta (Qian et al., 2007) for 58 proteins. To evaluate our potential in a more realistic setting, for each protein we used only the 100 low-quality decoys in the set, excluding the high-quality decoys. The average GDT score of the best decoys is about 60. As shown in Table 1, our EPAD, which currently considers only backbone atoms, correctly identifies 34 native structures with the lowest Z score (−2.46), while two full-atom potentials DOPE and DFIRE can identify only 21 natives. EPAD also exceeds DFIRE and DOPE in terms of the average ranking of the native structures (15.70 vs. 23.71 and 21.59, respectively). In terms of the average

Table 2. Performance of EPAD and Several Popular Statistical Potentials on the Rosetta Decoy Sets When Only C_α Atoms Are Considered

	EPAD	DOPE	DFIRE	MyDope	OPUS
No. of natives identified	**33**	11	12	10	6
Ranking of native	**15.8**	18. 7	30.7	21.7	55.3
First-ranked GDT score	**51.2**	47.0	47.8	48.2	45.9
Pearson CC	**−0.40**	−0.24	−0.20	−0.21	−0.15
Z score	**−2.45**	−1.51	−0.66	−1.23	0.25

Numbers in bold indicate the best performance. MyDope is a recompiled DOPE using the EPAD training data.

Table 3. Performance of EPAD and Several Popular Statistical Potentials on the Template-Based Models

	InHouse (%)	Set1 (%)	Set2 (%)	Set3 (%)	Set4 (%)
EPAD	**1,903 (53)**	**617 (54)**	**178 (54)**	**514 (53)**	**143 (54)**
DOPE	900 (25)	288 (25)	82 (25)	252 (26)	74 (28)
DFIRE	936 (26)	286 (25)	86 (26)	253 (26)	74 (28)
OPUS	900 (25)	289 (25)	73 (22)	251 (26)	69 (26)
RW	762 (21)	248 (22)	68 (21)	218 (23)	60 (22)

Only EPAD is a backbone-based potential, while the others are full-atom potentials. Data indicate the number (and percentage) of correctly identified models (i.e., models with the lowest energy value and the best GDT score). Bold numbers indicate the best performance.

per-target Pearson CC between the energy and GDT, EPAD (−0.42) is significantly better than DOPE (−0.32) and DFIRE (−0.25). EPAD also exceeds RW by all the five metrics.

EPAD compares favorably to OPUS-PSP, a full-atom statistical potential. OPUS can correctly identify many more native structures than EPAD, but it has a very low Pearson CC to decoy quality, which indicates that OPUS-PSP may not be good for ab initio folding. Since EPAD does not contain side chain atoms, we simply built a full-atom potential EPAD2 by linearly combining EPAD with the side chain component in OPUS-PSP (with equal weight). EPAD2 significantly outperformed DOPE, DFIRE, RW, and OPUS-PSP by all the five metrics. EPAD2 greatly outperformed EPAD in correctly recognizing the native structures, which may imply that side chain information is very helpful for the identification of the native structures. This trend was also observed on other data sets (e.g., I-TASSER and CASP5-8).

Table 2 compares the performance of several statistical potentials when only C_α atoms are considered in scoring a decoy. Again, EPAD significantly outperformed DOPE, DFIRE, and OPUS. EPAD even performed as well as the full-atom potentials DOPE, DFIRE, and RW. To exclude the impact of the different data sets used to build EPAD and DOPE, we rebuilt a DOPE using the EPAD training data, denoted as MyDope. MyDope performed slightly worse than DOPE, possibly because we did not fine-tune MyDope. However, EPAD performed significantly better than both DOPE and MyDope. This indicates that EPAD outperforms DOPE not because of the training data but because of the novel methodology.

Performance on Template-Based Models

To examine the performance of EPAD on template-based models, we constructed a large set of 3,600 protein pairs, denoted as InHouse, with the following properties: (1) Any two proteins in a pair share less than 25% sequence identity; (2) all protein classes (i.e., alpha, beta, and alpha-beta proteins) are covered by this set; (3) the protein lengths are widely distributed; (4) the structure similarity of two proteins in a pair, measured by TM score, ranges from 0.5 to 0.8 and is almost uniformly distributed; and (5) within each protein pair, one protein is designated as the target protein and the other as the template protein. Any two target proteins share less than 25% sequence identity. We generated four different target-template alignments for each protein pair using our in-house protein structure alignment program DeepAlign and our two threading programs BoostThreader (Peng and Xu, 2009) and CNFpred, as well as the popular profile alignment program HHpred (Söding, 2005). CNFpred is an improved version of BoostThreader, replacing regression trees in the latter with neural networks and making use of more protein features. Then we used MODELER (Eswar et al., 2006) to build four different three-dimensional (3D) models for each target protein based upon the four alignments, respectively. MODELER also builds models for the unaligned regions. To remove overlap with the training data, we constructed four subsets, Set1, Set2, Set3, and Set4 of InHouse as follows. Set1 contains no target proteins sharing >25% sequence identity with the EPAD training proteins. Set2 is a subset of Set1, containing no target proteins sharing >25% sequence identity with the EPAD validation proteins. Set3 contains no target proteins with a BLAST E value < 0.01 with the EPAD training proteins. Set4 is a subset of Set3, containing no target proteins with a BLAST E value < 0.01 with the EPAD training and proteins. In total, Set1, Set2, Set3, and Set4 contain 1139, 331, 965, and 266 protein pairs, respectively.

Table 3 lists the performance of several statistical potentials in identifying the 3D models with the best GDT score in the five data sets: InHouse, Set1, Set2, Set3, and Set4. As shown in Table 3, EPAD is able to recognize many more the best template-based models than the others, which are no better than random guess. Furthermore, EPAD has similar performance on the five sets, which confirms that our PNN model is not overtrained. For over 95% of protein pairs in InHouse, the 3D models built from the structure alignments have the best GDT score. This implies that, except EPAD, the other potentials are not able to differentiate structure alignments from threading-generated alignments.

Performance on the CASP9 Models

To further examine the performance of EPAD, we compiled a test set from the CASP9 models submitted by the top 18 servers. We excluded the CASP9 targets with many domains since some servers do not place the models of all the domains in a single coordinate system. These 18 servers are BAKER-ROSETTASERVER (Raman et al., 2009), chunk-TASSER (Zhou et al., 2009), chuo-fams (Kanou et al., 2009), CLEF-Server (Shao et al., 2011), FAMSD (Kanou et al., 2009), gws (Joo et al., 2009), HHpredA (Hildebrand et al., 2009), Jiang_Assembly (Hu et al., 2011), MULTICOM-CLUSTER (Tegge et al., 2009), MULTICOM-NOVEL (Tegge et al., 2009), Pcomb (Larsson et al., 2008), Phyre2 (Kelley and Sternberg, 2009), pro-sp3-TASSER (Zhou and Skolnick, 2009), QUARK (Xu et al., 2011), RaptorX (Peng

Table 4. Performance of Statistical Potentials with Respect to the Hardness of the CASP9 Targets

	GDT of the First-Ranked Models				Correlation Coefficient			
	EPAD	DP	DF	RW	EPAD	DP	DF	RW
<30	**27.4**	23.1	24.1	25.8	**0.44**	0.28	0.23	0.33
30–50	42.0	40.0	40.6	**42.7**	**0.31**	0.26	0.24	0.27
50–70	**64.4**	61.6	61.5	63.4	**0.37**	0.24	0.22	0.27
>70	**80.0**	77.1	77.1	77.1	**0.35**	0.26	0.26	0.26

To save space, DOPE and DFIRE are denoted as DP and DF, respectively, and we also omitted the negative sign of the correlation coefficient. The first column indicates the hardness of the targets, judged by the average GDT score of all the models of the target.

and Xu, 2011), Seok-server (Lee et al., 2010), Zhang-Server (Xu et al., 2011), and ZHOU-SPARKS-X (Yang et al., 2011). We did not include the models from RaptorX-MSA, RaptorX-Boost, HHpredB, HHpredC, MULTICOM-REFINE, MULTICOM-CONSTRUCT, and Jiang_THREADER since they are not very different from some of the 18 servers. In summary, this CASP9 data set contains the first models submitted by 18 servers for 92 targets. This set is very challenging for any energy potentials because the models submitted by these top servers have similar quality, especially for those not-so-hard targets. The first-ranked models by EPAD, DOPE, DFIRE, and RW have average GDT scores of 58.6, 55.7, 56.0, and 57.4, respectively. The average Pearson CC (between GDT scores and energy values) for EPAD is −0.364, which is significantly better than that for DOPE (−0.25), DFIRE (−0.23), and RW (−0.28). Note that RW parameters are fine-tuned using the CASP8 and CASP9 models while EPAD, DOPE, and DFIRE are independent of any decoy sets. In addition, EPAD is only a backbone-based potential while the other three are full-atom potentials.

Table 4 shows the performance of EPAD, DOPE, DFIRE, and RW with respect to the hardness of the targets, which is judged based upon the average GDT score of all the models of this target. We divide the targets into four groups according to the average GDT score: <30, 30–50, 50–70, and >70, respectively. EPAD performs very well across all difficulty levels and has a particularly good correlation coefficient for the targets with average GDT scores less than 30. Even for easy targets EPAD also outperforms the others although it is believed that sequence profiles are not very effective in dealing with easy targets. The only exception is that EPAD has a worse average GDT score of the first-ranked models than RW for the targets with an average GDT score between 30 and 50. This is because RW performs exceptionally well on a single target T0576. The best model identified by RW has a GDT score of 53.3, while EPAD, DOPE, and DFIRE can only identify a model with a GDT score of 17.0.

Performance on the Decoy 'R' Us Data Set

The set is taken from http://dd.compbio.washington.edu/, containing decoys for some very small proteins. In terms of the average rank of the native structures, EPAD significantly exceeds the others, but EPAD correctly identifies slightly fewer native structures than DOPE and OPUS_PSP, in part because EPAD does not include side chain atoms. See Supplemental Data for the details.

Performance on the I-TASSER Data Set

This set contains decoys for 56 proteins generated by I-TASSER (http://zhanglab.ccmb.med.umich.edu/). The average TM score of the decoys in this set ranges from 0.346 to 0.678. EPAD outperforms DFIRE and DOPE by five measures. EPAD is slightly better than RW in terms of the first-ranked TM score and the correlation but slightly worse than RW in terms of the Z score of the natives. EPAD2 (i.e., the combination of the OPUS-PSP side chain potential and EPAD) can obtain a much better Z score for the natives, although the correlation is slightly decreased. This is consistent with what was observed on the Rosetta set. See Supplemental Data (Table S2) for the details.

Performance on the CASP5-8 Data Set

See Rykunov and Fiser (2010) for the detailed account of this data set. EPAD is only worse than QMEAN6, RW, and RWplus in ranking the best models in the absence of the native structures. When the native structures are included, EPAD does not perform as well as when the native structures are not included. EPAD2 outperforms all the others in terms of the average ranking of the best models in the absence of the native structures or the average ranking of the native structures. EPAD2 also performs very well in terms of the number of correctly identified models (or native structures). These results may further indicate that side chain information is needed for the accurate identification of the native structures. See Supplemental Data (Table S3) for the details.

Is Our PNN Model Overtrained?

Our PNN model has 60,000–70,000 parameters to be trained. A natural question to ask is whether our PNN model is biased toward some specific patterns in the training data. Can our PNN model be generalized well to proteins of novel folds or sequences? According to our experimental results on contact prediction (see "Window size and the number of neurons" in the Supplemental Experimental Procedures), our PNN model is not overtrained. In this experiment, we used a training set built before CASP8 started, which is unlikely to have folds and sequence profiles similar to those in our test set (i.e., the CASP8 and CASP9 free modeling targets). Experimental results indicate that our PNN method compares favorably to the best CASP8 and CASP9 server predictors, which implies that our PNN model is not biased toward the training data. Note that our PNN model for statistical potential uses exactly the same architecture (two hidden layers with 100 and 40 hidden neurons, respectively) as our PNN model for contact prediction. Considering that much more training data (~73 millions of residue pairs) is used for the derivation of our statistical potential than for contact prediction, it is less likely that our PNN model for statistical potential is biased toward some specific patterns in the training set. The result in Table 3 further confirms this. We use the 25% sequence identity or an E value of 0.01 as the cutoff to exclude proteins in InHouse with similar sequences to the training set and generate two subsets, Set2 and Set4. Even if Set2 (Set4) contains some sequence profiles similar to the training set, the similarity between the whole InHouse set and the training set is still much larger than that between Set2 (Set4) and the training set, but the performance on the whole InHouse set is even slightly worse than that on Set2 (Set4).

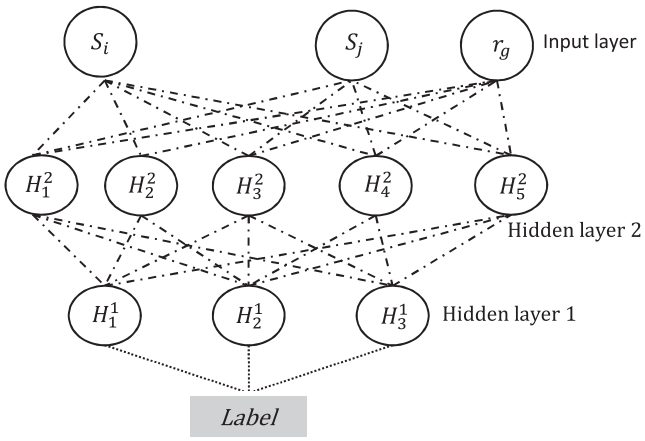

Figure 2. An Example Probabilistic Neural Network, in which S_i and S_j Are the Sequence Profile Contexts Centered at the i^{th} and j^{th} Residues, Respectively

H_q^1 and H_p^2 are the neurons in the first and second hidden layers.

Conclusions

This article presents a protein-specific and position-specific knowledge-based statistical potential, EPAD, for protein structure and functional study. EPAD has different energy profiles for two atoms of given types, depending on the protein under consideration and the sequence profile contexts of the residues containing them, while other potentials have the same energy profile for a given atom pair across all proteins. We achieve this by parameterizing EPAD using evolutionary information and radius of gyration of the protein under consideration in addition to atom types, which enables us to obtain a much more accurate statistical potential.

This article also makes a technical contribution to estimating the observed atomic interacting probability by introducing a probabilistic neural network to calculate the interatom distance probability distribution from sequence profiles and the radius of gyration. This is very different from the simple counting method widely used to derive the position-independent statistical potentials such as DOPE and DFIRE. The simple counting method does not work for our potential simply because there is not enough number of solved protein structures in the PDB for reliable counting of sequence profile contexts.

Experimental results indicate that EPAD significantly outperforms several popular higher resolution full-atom potentials in several decoy discrimination tests even if only backbone atoms are considered in EPAD. If we combine EPAD with the side chain component in OPUS-PSP, we can achieve much better decoy discrimination performance especially in the presence of native structures. As opposed to the RW potential and many others, EPAD is not trained by any decoys so, in principle, it is not restricted to any decoy generation method. Currently, EPAD uses only 1 Å resolution for the C_α-C_α distance discretization. We will further improve EPAD by using a 0.5 Å resolution, but this will take a very long time to train a neural network model for accurate estimation of the extremely unbalanced distance probability distribution.

We will continue to extend our statistical potential as follows. Currently, EPAD considers only backbone atoms and is also orientation independent. In the future, we will extend it to side chain atoms and also make it orientation dependent. Second, in estimating the distance probability distribution of two positions, we use only sequence profile contexts relevant to only these two positions. We shall also use information in the sequence segment connecting the two residues, which contains important information in determining the relative orientation of the two residues. Third, we may also estimate the distance probability distribution more accurately by adding some physical constraints. For example, given any three atoms in a protein, their pairwise distances must satisfy the triangle inequality. Furthermore, for any three residues that are close to one another along the primary sequence, their C_α distances are also subject to the restriction of local atomic interaction. If we assume that there is a contact between two residues if their C_α or C_β atoms are within 8 Å, then the number of contacts for any given residue is limited by a constant (~13) due to geometric restraint. By enforcing these constraints, we shall be able to estimate the interatom distance probability distribution much more accurately and, thus, design a much better statistical potential.

EXPERIMENTAL PROCEDURES

Estimating Pairwise C_α Distance Distribution Using the PNN

We discretize all the C_α-C_α distances into 13 bins (3–~4 Å, 4–5 Å, 5–6 Å, . . . , 14–15 Å, and >15 Å). Given a protein and its k^{th} residue pair of two residues i and j, let d_k denote the bin into which the distance of the k^{th} residue pair falls, and let x_k denote the position-specific feature vector, which contains sequence profile contexts S_i and S_j centered at the two residues i and j under consideration and the estimated radius of gyration of the protein under consideration.

Let S denote the sequence profile of the protein under consideration. It is generated by running PSI-BLAST on the NR database with, at most, eight iterations and an E value of 0.001. S is a position-specific scoring matrix with dimension $20 \times N$, where N is the sequence length. Each column in S is a vector of 20 elements, containing the mutation potential to the 20 amino acids at the corresponding sequence position. The sequence profile context of the residue at sequence position i is a 20×15 submatrix of S, consisting of 15 columns $i-7$, $i-6$, . . . i, $i+1$, . . . $i+7$. In case that one column does not exist in S (when $i \leq 7$ or $i+7 > N$), the zero vector is used.

We always use $r_g = 2.2N^{0.38}$ to estimate the radius of gyration for one protein where N is the protein sequence length. That is, r_g is independent of any 3D models including the native structure. We do not use r_g specific to a decoy because our training set does not contain any decoys. We do not use r_g calculated from the native structures either, because in the realistic settings they are unavailable.

Let $p_\theta(d_k|x_k)$ be the probability of the distance label d_k conditioned on the feature vector x_k. Meanwhile, θ is the model parameter vector. We estimate $p_\theta(d_k|x_k)$ as follows:

$$p_\theta(d_k|x_k) = \frac{\exp(\phi_\theta(x_k, d_k))}{Z_\theta(x_k)}, \quad (3)$$

where $Z_\theta(x_k) = \sum_d \exp(\phi_\theta(x_k, d))$ is the partition function and $\phi_\theta(x, d)$ is a two-layer neural network. Figure 2 shows an example of the neural network with three and five neurons in the first and second hidden layers, respectively. Each neuron represents a sigmoid function $h(x) = 1/(1+\exp(x))$. Therefore, we have

$$\phi_\theta(x_k, d_k) = \sum_{g_1=1}^{G_1} \theta^0_{d_k, g_1} h\left(\sum_{g_2=1}^{G_2} \theta^1_{g_1, g_2} h\left(<\theta^2_{g_2}, x_k> \right) \right), \quad (4)$$

where G_1 and G_2 are the number of gates in the two hidden layers, <.,.> denotes the inner product of two vectors, $\theta^2_{g_2}$ is the weight vector of the g_2^{th} neuron (also known as gate) in the second layer; $\theta^1_{g_1, g_2}$ is the weight connecting

the g_2^{th} neuron in the second layer to the g_1^{th} neuron in the first layer; and $\theta^0_{d_i,g_1}$ is the weight connecting the g_1^{th} neuron in the first layer to the label d_k.

In the implementation, our neural network consists of two hidden layers. The first hidden layer (i.e., the layer connecting to the input layer) contains 100 neurons, and the second hidden layer (i.e., the layer connecting to the output layer) has 40 neurons. This neural network is similar to what is used by the Zhou group (Xue et al., 2009) for interresidue contact prediction, which uses 100 and 30 neurons in the two hidden layers, respectively. The Zhou group has shown that using two hidden layers can obtain slightly better performance than using a single hidden layer. The input layer of our network has about 600 features, so in total, our neural network has between 60,000 and 70,000 parameters to be trained.

Model Parameter Training

We use the maximum likelihood method to train the model parameter θ and to determine the window size and the number of neurons in each hidden layer by maximizing the occurring probability of the native C_α-C_α distance in a set of training proteins. Given a training protein t with solved experimental structure, let D^t denote the set of pairwise residue distances and X^t denote the set of all feature vectors. By assuming any two residue pairs to be independent of one another, we have

$$p_\theta(D^t|X^t) = \prod_{k=1}^{m_t} p_\theta(d_k^t|x_k^t), \tag{5}$$

where m_t is the number of residue pairs in the protein t.

Given T training proteins, we need to maximize $\prod_{t=1}^{T} p_\theta(D^t|X^t)$, which is equivalent to the following optimization problem.

$$\min_\theta \sum_{t=1}^{T} -\log p_\theta(D^t|X^t) + \lambda||\theta||_2^2 = \min_\theta \sum_{t=1}^{T}\sum_{i=1}^{m^t}\left(-\log Z_\theta(x_k^t) + \phi_\theta(x_k^t, d_k^t)\right) + \lambda||\theta||_2^2. \tag{6}$$

Meanwhile, $\lambda||\theta||_2^2$ is a L_2-norm regularization item to avoid overfitting and λ is a hyperparameter to be determined. This optimization problem can be solved by the limited-memory BFGS method (Liu and Nocedal, 1989).

It is very challenging to solve this nonconvex optimization problem due to the huge amount of training data. We generated an initial solution randomly and then ran the training algorithm on a supercomputer for about a couple of weeks. Our training algorithm terminated when the probability of either the training set or the validation set did not improve any more. Note that all the model parameters are learned from the training set but not the validation set. The validation set, combined with the training set, is only used to determine when our training algorithm shall terminate. Our training algorithm usually terminates after 3,000 iterations. We also reran our training algorithm starting from nine initial solutions and did not observe explicit performance difference among these runs.

Training and Validation Data

We used the PDB25 set of the PISCES server (Wang and Dunbrack, 2003) early in 2011 as the training and validation data. Any two proteins in PDB25 share no more than 25% sequence identity. Such a set in total includes more than 6,000 proteins. We randomly chose about 5,000 proteins from this PDB25 set as the training and validation proteins and also made sure that they had no overlap (i.e., > 25% sequence identity) with the Rosetta set (Qian et al., 2007) and the Decoy 'R' Us set (Samudrala and Levitt, 2000). We randomly choose 3/4 of the 5,000 proteins as the training data and the remaining 1/4 as the validation data, which contain ~73 million training and ~19 million validation residue pairs, respectively. It is challenging to train our neural network model because (1) the number of training residue pairs is huge; and (2) the distance distribution is extremely unbalanced. As shown in Figure S3, 90% of residue pairs have C_α distance larger than 15 Å and only 1% of them have C_α distance less than 4 Å. It takes a couple of weeks to train a single neural network model using 1,296 CPUs on a Cray supercomputer.

Estimating Interatom Distance Distribution for Non-C_α Main Chain Atoms

We discretize the interatom distance of non-C_α atoms into 26 equal-width bins, each with 0.5 Å. Due to limited computation resources, instead of training neural network models for each pair of atom types, which will take months or even a year to finish, we use a different approach to estimate the pairwise distance probability distribution for non- C_α main chain atoms. In particular, we calculate the interatom distance probability distribution for non-C_α main chain atoms conditioned upon C_α-C_α distance probability distribution. Let $P_{\alpha\alpha}(d_{\alpha\alpha}|S_i, S_j, r_g)$ denote the C_α-C_α distance probability distribution for residues i and j, which can be estimated by our probabilistic neural network. Let a and b denote the amino acid types of the residues at i and j, respectively. For the purpose of simplicity, we use N and O atoms as an example to show how to calculate the observed atomic interacting probability. Let $P(d|N, O, S_i, S_j, r_g)$ denote the distance probability distribution for the nitrogen atom in residue i and the oxygen atom in residue j. We calculate $P(d|N, O, S_i, S_j, r_g)$ as follows:

$$P(d|N, O, S_i, S_j, r_g) = \sum_{d_{\alpha\alpha}} P_{NO}^{a,b}(d|d_{\alpha\alpha}) P_{\alpha\alpha}(d_{\alpha\alpha}|S_i, S_j, r_g), \tag{7}$$

where $P_{NO}^{a,b}(d|d_{\alpha\alpha})$ is the conditional distance probability distribution for atom N in amino acid a and O in amino acid b when the C_α distance of these two amino acids is $d_{\alpha\alpha}$. Since $P_{NO}^{a,b}(d|d_{\alpha\alpha})$ is position independent, it can be estimated by simple counting.

Window Size and the Number of Neurons in the Hidden Layers

The window size for a sequence profile context and the number of neurons in the hidden layers are important hyperparameters of our probabilistic neural network. Because it is time consuming to train even a single neural network model for the estimation of distance probability distribution, we determined these hyperparameters by training a neural network for interresidue contact prediction, which obtains the best performance when the window size is 15 and the numbers of neurons in the first and second hidden layers are 40 and 100, respectively. Details are shown in Supplemental Experimental Procedures. The window size that we used is consistent with what was used by the Zhang group (Wu and Zhang, 2008a), and the numbers of hidden neurons are not very different from what was used by the Zhou group (Xue et al., 2009).

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, three tables, Supplemental Data, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.str.2012.04.003.

ACKNOWLEDGMENTS

We thank Jianzhu Ma and Sheng Wang for helping generate the in-house template-based models. This work is supported by National Institutes of Health Grant R01GM0897532 and National Science Foundation Grant DBI-0960390. We are grateful to the University of Chicago Beagle team, TeraGrid, and Canadian SHARCNet for their support of computational resources.

Received: January 14, 2012
Revised: April 9, 2012
Accepted: April 10, 2012
Published online: May 17, 2012

REFERENCES

Bauer, A., and Beyer, A. (1994). An improved pair potential to recognize native protein folds. Proteins *18*, 254–261.

Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D.E., Meiler, J., Misura, K.M.S., and Baker, D. (2005). Free modeling with Rosetta in CASP6. Proteins *61* (*Suppl 7*), 128–134.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009). CHARMM: the biomolecular simulation program. J. Comput. Chem. *30*, 1545–1614.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins *21*, 167–195.

Bystroff, C., Thorsson, V., and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J. Mol. Biol. *301*, 173–190.

Casari, G., and Sippl, M.J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. J. Mol. Biol. *224*, 725–732.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J. Comput. Chem. *26*, 1668–1688.

Dill, K.A. (1985). Theory for the folding and stability of globular proteins. Biochemistry *24*, 1501–1509.

Dill, K.A. (1997). Additivity principles in biochemistry. J. Biol. Chem. *272*, 701–704.

Dobson, C.M., Sali, A., and Karplus, M. (1998). Protein folding: A perspective from theory and experiment. Angew. Chem. Int. Ed. Eng. *37*, 868–893.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling with MODELLER. Curr. Protoc. Bioinformatics *5* (*Suppl 15*), 5.6.1–5.6.30.

Gatchell, D.W., Dennis, S., and Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins *41*, 518–534.

Gilis, D., and Rooman, M. (1996). Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. J. Mol. Biol. *257*, 1112–1126.

Gilis, D., and Rooman, M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J. Mol. Biol. *272*, 276–290.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. *216*, 167–180.

Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009). Fast and accurate automatic structure prediction with HHpred. Proteins *77* (*Suppl 9*), 128–132.

Hu, Y., Dong, X., Wu, A., Cao, Y., Tian, L., and Jiang, T. (2011). Incorporation of local structural preference potential improves fold recognition. PLoS ONE *6*, e17215.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

Jones, D.T., and Thornton, J.M. (1996). Potential energy functions for threading. Curr. Opin. Struct. Biol. *6*, 210–216.

Jones, T.A., and Thirup, S. (1986). Using known substructures in protein model building and crystallography. EMBO J. *5*, 819–822.

Joo, K., Lee, J., Seo, J.H., Lee, K., Kim, B.G., and Lee, J. (2009). All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. Proteins *75*, 1010–1023.

Kanou, K., Iwadate, M., Hirata, T., Terashi, G., Umeyama, H., and Takeda-Shitaka, M. (2009). FAMSD: A powerful protein modeling platform that combines alignment methods, homology modeling, 3D structure quality estimation and molecular dynamics. Chem. Pharm. Bull. (Tokyo) *57*, 1335–1342.

Kelley, L.A., and Sternberg, M.J.E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. *4*, 363–371.

Kihara, D., Lu, H., Kolinski, A., and Skolnick, J. (2001). TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc. Natl. Acad. Sci. USA *98*, 10125–10130.

Kortemme, T., and Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. Proc. Natl. Acad. Sci. USA *99*, 14116–14121.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., and Berman, H.M. (2006). The RCSB PDB information portal for structural genomics. Nucleic Acids Res. *34* (Database issue), D302–D305.

Larsson, P., Wallner, B., Lindahl, E., and Elofsson, A. (2008). Using multiple templates to improve quality of homology models in automated homology modeling. Protein Sci. *17*, 990–1002.

Laurie, A.T.R., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics *21*, 1908–1916.

Lee, J., Lee, D., Park, H., Coutsias, E.A., and Seok, C. (2010). Protein loop modeling by using fragment assembly and analytical loop closure. Proteins *78*, 3428–3436.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. *226*, 507–533.

Liu, D., and Nocedal, J. (1989). On the limited memory Bfgs method for large-scale optimization. Math. Program. *45*, 503–528.

Lu, H., and Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins *44*, 223–232.

Lu, M., Dousis, A.D., and Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J. Mol. Biol. *376*, 288–301.

Maiorov, V.N., and Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. *227*, 876–888.

Miyazawa, S., and Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal-structures: quasi-chemical approximation. Macromolecules *18*, 534–552.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.

Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. (2000). Combination of threading potentials and sequence profiles improves fold recognition. J. Mol. Biol. *296*, 1319–1331.

Panjkovich, A., Melo, F., and Marti-Renom, M.A. (2008). Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs. Genome Biol. *9*, R68.

Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. *36*, 2295–2300.

Peng, J., and Xu, J. (2009). Boosting protein threading accuracy. Res. Comput. Mol. Biol. *5541*, 31–45.

Peng, J., and Xu, J. (2010). Low-homology protein threading. Bioinformatics *26*, i294–i300.

Peng, J., and Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins *79* (*Suppl 10*), 161–171.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature *450*, 259–264.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J.M., Kim, D., Kellogg, E., DiMaio, F., Lange, O., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins *77* (*Suppl 9*), 89–99.

Reva, B.A., Finkelstein, A.V., Sanner, M.F., and Olson, A.J. (1997). Residue-residue mean-force potentials for protein structure recognition. Protein Eng. *10*, 865–876.

Rykunov, D., and Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics *11*, 128.

Samudrala, R., and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. *275*, 895–916.

Samudrala, R., and Levitt, M. (2000). Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci. *9*, 1399–1401.

Schuler, L.D., Daura, X., and Van Gunsteren, W.F. (2001). An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J. Comput. Chem. *22*, 1205–1218.

Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem. Rev. *106*, 1559–1588.

Shao, M.F., Wang, S., Wang, C., Yuan, X.Y., Li, S.C., Zheng, W.M., and Bu, D.B. (2011). Incorporating Ab Initio energy into threading approaches for protein structure prediction. BMC Bioinformatics *12* (*Suppl 1*), S54.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci. *15*, 2507–2524.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins *34*, 82–95.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. *213*, 859–883.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins *17*, 355–362.

Sippl, M.J., and Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins *13*, 258–271.

Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. Curr. Opin. Struct. Biol. *16*, 166–171.

Skolnick, J., Kolinski, A., and Ortiz, A. (2000). Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins *38*, 3–16.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951–960.

Specht, D.F. (1990). Probabilistic neural networks. Neural Netw. *3*, 109–118.

Tanaka, S., and Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules *9*, 945–950.

Tegge, A.N., Wang, Z., Eickholt, J., and Cheng, J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res. *37* (Web Server issue), W515–W518.

Vendruscolo, M., Najmanovich, R., and Domany, E. (2000). Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins *38*, 134–148.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

Wang, Z., Zhao, F., Peng, J., and Xu, J. (2011). Protein 8-class secondary structure prediction using conditional neural fields. Proteomics *11*, 3786–3792.

Wu, S., and Zhang, Y. (2008a). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics *24*, 924–931.

Wu, S., and Zhang, Y. (2008b). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins *72*, 547–556.

Wu, S., Skolnick, J., and Zhang, Y. (2007a). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 5–17.

Wu, Y., Lu, M., Chen, M., Li, J., and Ma, J. (2007b). OPUS-Ca: a knowledge-based potential function requiring only Calpha positions. Protein Sci. *16*, 1449–1463.

Xu, D., Zhang, J., Roy, A., and Zhang, Y. (2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins *79* (*Suppl 10*), 147–160.

Xu, J. (2005). Fold recognition by predicted alignment accuracy. IEEE/ACM Trans. Comput. Biol. Bioinformatics *2*, 157–165.

Xue, B., Faraggi, E., and Zhou, Y. (2009). Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins *76*, 176–183.

Yang, Y., Faraggi, E., Zhao, H., and Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics *27*, 2076–2082.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. *31*, 3370–3374.

Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. Proteins, 22–29.

Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. (2001). Processing and evaluation of predictions in CASP4. Proteins (*Suppl 5*), 13–21.

Zhang, C., Vasmatzis, G., Cornette, J.L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. J. Mol. Biol. *267*, 707–726.

Zhang, J., and Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS ONE *5*, e15386.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.

Zhao, F., Li, S., Sterner, B.W., and Xu, J. (2008). Discriminative learning for protein conformation sampling. Proteins *73*, 228–240.

Zhao, F., Peng, J., and Xu, J. (2010). Fragment-free approach to protein folding using conditional neural fields. Bioinformatics *26*, i310–i317.

Zhou, H., and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. *11*, 2714–2726.

Zhou, H., and Skolnick, J. (2009). Protein structure prediction by pro-Sp3-TASSER. Biophys. J. *96*, 2119–2127.

Zhou, H., Pandit, S.B., and Skolnick, J. (2009). Performance of the Pro-sp3-TASSER server in CASP8. Proteins *77* (*Suppl 9*), 123–127.

