Biophysical Journal Volume 103 November 2012 1950–1959

Using the Unfolded State as the Reference State Improves the Performance of Statistical Potentials

Yufeng Liu and Haipeng Gong*
Ministry of Education Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China

ABSTRACT Distance-dependent statistical potentials are an important class of energy functions extensively used in modeling protein structures and energetics. These potentials are obtained by statistically analyzing the proximity of atoms in all combinatorial amino-acid pairs in proteins with known structures. In model evaluation, the statistical potential is usually subtracted by the value of a reference state for better selectivity. An ideal reference state should include the general chemical properties of polypeptide chains so that only the unique factors stabilizing the native structures are retained after calibrating on reference state. However, reference states available as of this writing rarely model specific chemical constraints of peptide bonds and therefore poorly reflect the behavior of polypeptide chains. In this work, we proposed a statistical potential based on unfolded state ensemble (SPOUSE), where the reference state is summarized from the unfolded state ensembles of proteins produced according to the statistical coil model. Due to its better representation of the features of polypeptides, SPOUSE outperforms three of the most widely used distance-dependent potentials not only in native conformation identification, but also in the selection of close-to-native models and correlation coefficients between energy and model error. Furthermore, SPOUSE shows promising possibility of further improvement by integration with the orientation-dependent side-chain potentials.

INTRODUCTION

With the rapid development of genomics, especially the success of Human Genome Project, hundreds of thousands protein-encoding sequences have been deposited into the gene database. As a contrast, only ~73,000 of these sequenced proteins have their structures determined (1), constituting <0.4% of the >20,000,000 sequence ensemble (2). This huge gap between protein sequences and structures is still broadening, and can only be filled by computational modeling and structural prediction in the lack of breakthrough in structure-determining techniques (3). In a typical structural prediction algorithm, numerous structural models are sampled in silico and evaluated by an energy function, and finally the one with lowest energy is frequently chosen as the best model, under the assumption that the native conformation is energetically more favorable than all other ones. Therefore, a precise energy function (or potential) is the prerequisite for accurate protein structural prediction.

The potential functions can be roughly divided into two major categories: physical potentials and statistical potentials (4). Physical potentials, including CHARMM (5), AMBER (6), GROMOS (7), and OPLS (8), are summarized from physical laws and have been widely adopted in molecular dynamics simulations. However, they are not only time-consuming (9), but also weak in selecting good models generated by real-time protein structural prediction algorithms (10). More importantly, physical force fields neglect the entropy effect—a factor that must be considered in protein folding and free energy calculation. Statistical potentials, on the other hand, are derived from the high-resolution structures deposited in the Protein DataBank (PDB) (1). Due to their high efficiencies in selecting the close-to-native conformations, they have been widely used in the protein structure prediction, *ab initio* folding, model assessment, etc (9).

Based on the interaction type, statistical potentials can be further classified into several categories, including distance-dependent potentials (9–20), contact-based potentials (21–24), and numerous other interaction-type based potentials. One of the first reported statistical potentials was a residue-level contact-based potential that incorporated contact frequencies over short-range, medium-range, and long-range (21). Many statistical potentials have been developed since then, among which the distance-dependent potential is the most commonly used (9).

To date, all distance-dependent statistical potentials are based on two assumptions:

1. the native structure corresponds to the lowest energy model; and
2. according to Boltzmann assumption, energy is proportional to the negative logarithm of the state probability, which can be expressed as

$$E = -RT\ln(P), \tag{1}$$

where R is the ideal gas constant, T is the temperature, and P is the state probability.

To rule out the nonspecific chemical features of polypeptides, a proper reference state is often integrated, which makes Eq. 1 as

$$E = -RT\ln\left(\frac{P_{obs}}{P_{exp}}\right), \tag{2}$$

where P_{obs} is the observed probability derived from high-resolution crystal structures, and P_{exp} is the expected

Submitted May 6, 2012, and accepted for publication September 19, 2012.
*Correspondence: hgong@tsinghua.edu.cn
Editor: Doug Barrick.

0006-3495/12/11/1950/10 $2.00
http://dx.doi.org/10.1016/j.bpj.2012.09.023

probability of the reference state. Because the universal crystal structure database is used by all distance-dependent potentials in estimating P_{obs}, the major difference comes from the choice of reference state. The first distance-dependent potential, introduced in the 1990s by Sippl et al. (11), took the average of all atom pairs in the crystal structures as the reference state. In successive studies, several other potentials were built similarly by estimating P_{exp} from the PDB database, but the structural details of the native state had to be removed from the reference state, either by averaging all-atom pairs or by shuffling the atomic positions (10,13,14).

Besides statistical analysis through the database, the reference state could also be derived from physical modeling and theoretical deduction. Three distance-dependent potentials, DFIRE (15), DOPE (9), and random walk (RW) (19), were created sequentially, all of which are among the most popular energy functions used in structural modeling and prediction nowadays. DFIRE modeled the reference state as finite ideal gas, where the distance distribution of two randomly placed atoms follows the power law $(r^{\alpha}, \alpha = 1.61)$. The clean form of the reference state and the good performance in model evaluation render the popularity of DFIRE. Nevertheless, it is less effective in evaluating distantly placed atom pairs, because atomic pairwise distance in the reference state can increase unlimitedly so as to exceed the overall polypeptide chain length. DOPE modified the reference state to a homologous sphere isovolumic to the native protein to restrain the atomic distance, but it neglected the chain connectivity and volume of polypeptides. RW further improved the reference state by introducing the chain connectivity according to the random-walk model in the polymer theory. However, the chain volume is overlooked in the random-walk model. In addition, the polymer theory (and therefore the random-walk model) cannot completely describe the key properties of polypeptides, especially their tendencies to fold into unique three-dimensional native structures.

Theoretically, the best reference state should include properties universal to all polypeptide chains and exclude the properties of folded proteins so that only the unique attributes of the native conformations are retained in the statistical potential after normalization upon the reference state (Eq. 2). Unfortunately, none of the above-mentioned reference states fulfill this requirement completely. Intuitively, the unfolded state of a protein chain could be a good candidate, not only because it contains all chemical-linking information about the polypeptide chain, but also because nonbonded interactions are nearly absent in it. Moreover, this strategy agrees with the theoretical thinking of the protein folding process from statistical mechanics. For instance, when the unfolded state is taken as the reference, Eq. 2 automatically becomes the free energy change over the folding process.

The unfolded state is of key importance in protein folding research and has been extensively studied in the last decade (25,26). Proteins in unfolded state were thought to be structurally featureless random-coil polymers for a long time (27), because the radius of gyration predicted according to this model coincided with experimental data (28,29). Nevertheless, both experiments (30,31) and computational calculations (32,33) illustrated that considerable nativelike local topologies remain even in the extremely denatured proteins (26,34). In 2005, Jha et al. (35) proposed a statistical coil model, which reconciled the discrepancy between the radius-of-gyration scale and the residual nativelike topology. Furthermore, their model agreed well with the residual dipolar coupling data measured by NMR experiment under denaturing conditions. Fitzgerald et al. (4) calculated the P_{exp} from the unfolded state model of ubiquitin, and reported improvement of their potential in selecting reduced protein structures, in which all atoms beyond β-carbons are removed. Despite their success, derivation of the reference curve from a single protein impedes its application in longer protein chains.

Designed to facilitate structural prediction, the performance of statistical potentials ideally should be evaluated by their capability to promote the folding of protein structures in real-time molecular simulation, an approach hard to conduct due to the high computational expense. As an alternative, the statistical potentials are frequently tested using the decoy set, a set of structures with identical primary sequences to the native conformation. However, that method is challenged by the diversity in the protocols required to produce the decoys. During the submission of our study, Deng et al. systematically compared the performance of currently available statistical potentials using different decoy sets, and concluded "the performance of the potentials relies on the origin of decoy generations and no reference state can clearly outperform others in all decoy sets" (36).

In this work, we statistically reanalyzed the pairwise atomic distances in the unfolded state model of numerous protein chains, summarized an empirical formula to describe the dependence of this probability distribution on protein chain lengths, and took this formula as the reference state curve (P_{exp} in Eq. 2) to generate a novel, to our knowledge, all-atom distance-dependent statistical potential (SPOUSE). Testing upon decoy sets shows that SPOUSE is more powerful in native structure recognition and competitive in best model selection. Furthermore, it can be greatly improved by integrating an orientation-dependent side-chain potential term. The program SPOUSE is available at http://166.111.152.91/SPOUSE.html for free downloading.

THEORY

Mathematical background

According to Boltzmann assumption, protein potential is proportional to the negative logarithm of the probability of a given state as

$$E = -RT \cdot \ln(P(x_1, x_2, \ldots, x_N)), \quad (3)$$

where E is the total potential of a molecule, R is the ideal gas constant, T denotes the temperature, and $P(x_1,x_2,\ldots,x_N)$ is the probability of the state in which the N atoms are described by their coordinates $x_1,x_2,\ldots,x_N$. Due to the high computational expense and the limited data in PDB, the N-dimensional joint probability is often reduced to the product of all pairwise probabilities in practice by neglecting the many-body interactions (9). Consequently, Eq. 3 is transformed to

$$E = -RT \sum_{i \neq j} \ln(P(xi, xj)), \quad (4)$$

where i and j refer to two interacting atoms. In other words, the total potential of a given protein conformation can be approximated by the sum of all pairwise potentials. For distance-dependent statistical potentials, Eq. 4 can be further deduced to the following formula when taking account of the reference state,

$$E = -RT \sum_{i \neq j} \ln\left(\frac{P_{obs}(r_{ij})}{P_{exp}(r_{ij})}\right), \quad (5)$$

where r_{ij} is the distance between atom i and j, and $P_{obs}(r_{ij})$ and $P_{exp}(r_{ij})$ are the observed and expected pairwise probability, respectively. In structural predictions, to significantly reduce the computational expense, usually only the atom pairs located within a certain cutoff value R_0 are considered in the energy estimation. Therefore, all statistical potentials actually force the energy at R_0 to zero and output the relative energy of a structure as

$$E = -RT \sum_{i \neq j} \ln\left(\frac{P_{obs}(r_{ij})/P_{obs}(R_0)}{P_{exp}(r_{ij})/P_{exp}(R_0)}\right). \quad (6)$$

Both observed and expected probabilities are thereby normalized by the respective probabilities at the cutoff distance.

Derivation of the observed probability

The observed pairwise probability of SPOUSE $P_{obs}(r_{ij})$ was calculated using selected high-resolution nonredundant crystal structures. The relative position between atom pairs was omitted, based on the report that symmetric potentials outperform their asymmetric counterparts (37).

Derivation of the expected probability

One-thousand unfolded models for protein chains of various lengths were produced to compute the expected probability. As shown in Fig. S3 in the Supporting Material, 1000 models are sufficient to capture the pairwise atomic distance distribution. The backbone and side-chain atoms were treated separately, because the former is spatially more constrained.

The pairwise distances between backbone atoms are all represented by distances between the α-carbons (CA) of the two corresponding residues for simplification in this work. Histogram of CA-CA distances from the unfolded state model of ubiquitin (see Fig. S1) indicates a rather noisy distribution within 10 Å and a much smoother distribution at longer distance. The discrete distribution at short distance mainly corresponds to the interaction between atoms residing in neighboring ($N1$) or alternating ($N2$) residues, respectively. Hence, the overall distribution is finally estimated by the weighted average of a local and nonlocal distribution, where the local one contains the $N1$ and $N2$ terms and the nonlocal one describes the distances between atoms separated by no less than two residues in the primary sequence (called $NL3$ here). For simplicity, the local distribution is further reduced to one single $N12$ term and represented by a Gaussian function. Finally, by weighted averaging of the local and nonlocal distribution (P_{12} and P_{NL3}), the pairwise expected probability for backbone atoms P_{exp} is formulated as

$$P_{exp} = w_{12}P_{12} + w_{NL3}P_{NL3}, \quad (7)$$

where w_{12} and w_{NL3} are the weights for the local and nonlocal distribution and could be estimated by counting the respective interacting pairs (e.g., $2N$-3 local residue pairs and $N(N$-1)/2-($2N$-3) nonlocal residue pairs) in a chain with N residues as

$$w_{12} = \frac{2N-3}{N(N-1)/2} \text{ and } w_{NL3} = 1 - w_{12}.$$

The local distribution (P_{12}) is universal for proteins of various sizes and is simply represented by a Gaussian distribution with the mean (5.05 Å) and standard deviation (1.29 Å) estimated from the $N12$ distances of CA-CA atom pairs in all unfolded state conformations. The standard deviation is later multiplied by a factor for the following two reasons:

1. CA-CA atom pairs in the unfolded state models are artificially too much confined, possibly because of the fixed bond-length and ignorance of *cis* peptide bonds in modeling (35). For instance, the $N1$ distance has an extremely small standard deviation (0.0004 Å), tremendously less than the actual value calculated from the PDB (0.02 Å).
2. The Gaussian distribution is obtained from CA-CA atom pairs and therefore should be flattened so as to represent all backbone atom pairs. The multiplication factor was tested among four possible choices (2,4,6,8) upon the

decoy sets *4state_reduced* (38) and *lmds* (39), and its value was finally set to "6" for the best performance.

However, the nonlocal distribution (P_{NL3}) should be dependent on the chain length. For a protein chain of a fixed length, the nonlocal distribution is found to follow the log-normal distribution

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(\ln x-\mu)^2}{2\sigma^2}\right], x>0$$

well. Therefore the *NL*3 histograms for all unfolded protein chains are fit with log-normal distribution and both parameters of log-normal distribution, μ and σ, show clear dependence on the chain length N (see Fig. S2). Fortunately these dependences can be well described by simple empirical formulas (linear relationship for μ and logarithm relationship for σ), as corroborated by the high-correlation coefficients in curve-fitting shown in Fig. S2. In other words, both μ and σ can be accurately predicted and then the nonlocal distribution P_{NL3} is known, once the number of residues in a chain is given. After integrating the P_{12} distribution using Eq. 7, the overall probability P_{exp} for backbone atoms can be easily derived.

All side-chain atoms beyond the β-carbon were absent in unfolded state ensembles and therefore their positions were predicted using SCWRL4 (40) according to the backbone topology. Though highly optimized, uncertainty may be introduced during the process. For this reason, chain length is neglected in treating these atoms and the distance distribution is obtained by counting over all side-chain atom pairs in all unfolded state models. The distance also follows a log-normal distribution, with μ and σ being 4.29 and 0.78, respectively. In the end, σ is slightly amplified to 0.84 for better performance tested in the decoy sets *4state_reduced* (38) and *lmds* (39).

MATERIALS AND METHODS

Crystal structure database

The nonredundant structures used to generate the observed probability were culled from PDB using the PISCES server (41). Only the structures determined by x-ray crystallography were preserved and all chains with missing internal residues were removed. The database contains 2724 chains with resolution <1.8 Å, *R*-factor <0.25, and pairwise sequence similarity <25%. Within the database, 101 chains are identified as having >60% sequence identity to the proteins in the test set (decoy set). After the removal of these homologous proteins, the final database contains 2623 chains.

Calculation of the observed probability

As with DOPE (9), SPOUSE contains 158 nonhydrogen residue-based atom types after the deletion of nine chemically equivalent atoms. Thus, a total number of 12,561 symmetric atom pairs are sampled. In practice, the probability is normalized according to the value at cutoff distance (see Eq. 6). Data beyond the cutoff distance is ignored. In this work, the cutoff distance is 15 Å except if otherwise mentioned. To obtain the distribution, the bin-width is set to 0.5 Å, the values at the bin midpoints are counted from the database, and all other values are estimated by linear interpolation.

Estimation of the expected probability

Thirteen nonhomologous chains (from 52 to 391 residues) were chosen (see Table S1 in the Supporting Material), and 1000 unfolded state models were generated for each of them by the Godzilla webserver (http://godzilla.uchicago.edu/cgi-bin/unfolded.cgi) (35). Missing heavy atoms of the side chains were replenished using SCWRL4 (40). The data fitting (in the Theory section) was achieved through *fitdistr*() functions in the R statistical package (http://www.r-project.org/) and the scientific PYTHON package (http://www.scipy.org/).

Potential evaluation and comparison

The performance of SPOUSE was tested in eight popularly used decoy sets (Table 1). The decoy sets *4state_reduced* (38), *lmds* (39), *fisa* (42), *fisa_casp3* (42), and *lattice_ssfit* (43,44), were downloaded from the Decoys 'R' Us database (45) (http://dd.compbio.washington.edu/). The comparative-modeling-based decoy set *molder* (46) was downloaded from the Sali lab (ftp://salilab.org/decoys/comp_models.tar.gz). Another two real-time-simulation-derived decoy sets *ROSETTA* (47) and *I-TASSER* (19) were obtained from the websites http://depts.washington.edu/bakerpg/decoys/rosetta_decoys_62proteins.tgz and http://zhanglab.ccmb.med.umich.edu/decoys/decoy2.html, respectively.

The three widely used distance-dependent potentials DFIRE, DOPE, and RW were tested for comparison. In addition, a statistical potential including orientation-dependent side chains, named GOAP (20), was downloaded from http://cssb.biology.gatech.edu/GOAP to evaluate possible improvements by integrating orientation-dependent features. For a strict comparison of the choice of reference state, DFIRE and RW potentials in this work were calculated in a similar way to SPOUSE, by replacing the SPOUSE reference curve with the theoretical formulas of DFIRE and RW taken from the respective published articles. The DOPE potential was calculated using the *assess_dope* function in MODELLER9v8 (48) due to the lack of detailed instruction on how to estimate the parameter *a* in the original article.

RESULTS

Comparison of the reference state

In Fig. S4, the probability distribution P_{exp} is calculated for a 76-residue chain and is overlaid on the top of the histogram of CA-CA distance (obtained from the unfolded state

TABLE 1 Native recognition test

Decoys	DFIRE	DOPE	RW	SPOUSE	Targets
fisa	3(70.5)	3(90.0)	3(72.3)	3(63.3)	4
fisa_casp3	5(1.0)	4(2.4)	5(1.0)	5(1.0)	5
lattice_ssfit	8(1.0)	8(1.0)	8(1.0)	8(1.0)	8
molder	19(6.5)	19(6.2)	19(6.6)	18(6.7)	20
ROSETTA	22(22.8)	22(24.6)	21(23.4)	27(19.2)	59
I-TASSER	51(4.0)	48(28.4)	52(2.1)	53(1.7)	56
Total(average)	108(13.1)	104(23.3)	108(12.7)	114(10.7)	152

Numbers outside the parentheses are the numbers of identified native conformations, and the ones in the parentheses are the average ranks of native structures within the decoy sets.

models of ubiquitin which also contains 76 residues). DFIRE and RW reference curves are also included for visual comparison. The SPOUSE empirical distribution curve agrees fairly well with the histogram, greatly exceeding both DFIRE and RW theoretical distributions. This indicates that the SPOUSE backbone reference state better captures the properties of polypeptide chains.

Both backbone and side-chain SPOUSE reference curves are created for a 100-residue chain, normalized by the value at $R_0 = 15$ Å, and plotted in Fig. 1. DFIRE and RW curves are shown again for comparison. DFIRE and RW reference curves are almost indistinguishable after normalization. RW and DOPE articles (9,19) showed that the DOPE curve is also very close to both DFIRE and RW curves after normalization. This explains the close performance of these three potentials listed in the literature, because limited progress has been made in capturing the essential characteristics of polypeptide chains despite the successive improvements in modeling the reference state. On the contrary, although the discrepancy between the SPOUSE backbone curve and the DFIRE and/or RW curves reduces significantly after normalization, great difference still exists, especially at distance <5 Å, where the SPOUSE backbone curve is significantly higher than the other models. The SPOUSE side-chain curve, however, is lower than the DFIRE and/or RW curves when the distance is <8 Å. According to Eq. 2, these changes will weaken the backbone-backbone interactions and strengthen the interactions including side-chain atoms at a short distance.

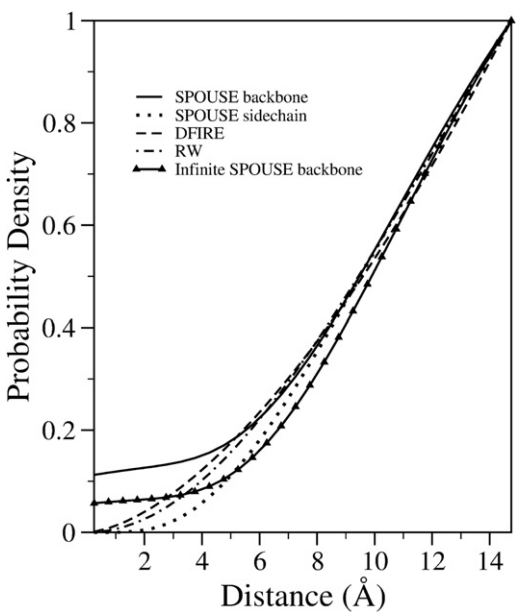

FIGURE 1 Comparison of reference states for different potentials. Five reference states, including SPOUSE backbone (*solid line*), SPOUSE side chain (*dotted line*), DFIRE (*dashed line*), RW (*dash-dotted line*), and infinite SPOUSE backbone (*triangle line*), which equally weights the separate terms as RW are drawn after normalized at cutoff distance for a 100-residue polypeptide.

Comparison of the energy function

The energy functions are plotted in Fig. S5 as examples for four different types of pairwise atomic interactions:

1. Backbone-backbone interaction.
2. Backbone-sidechain interaction.
3. Interactions between hydrophobic side-chain atoms.
4. Interactions between hydrophobic and hydrophilic side-chain atoms.

As expected from their indistinguishable reference curves, DFIRE and RW potentials are quite close to each other in all four cases.

In Fig. S5 *A*, both DFIRE and RW potentials experience the deepest trough at 2.2 Å, indicating extremely strong backbone-backbone interactions there. A simple analysis over crystal structures on the distribution of distance between atom pairs residing in neighboring residues (*N*1) and all other atom pairs (*NL*2) suggests that only *N*1 atom pairs are heavily populated in this region (2.2 Å). (see Fig. S6). Therefore, the deep trough at 2.2 Å is principally caused by the large population of *N*1 atom pairs, a group of atom pairs playing negligible roles in conformation selection, because *N*1 pairs make roughly equal energetic contribution in two conformations as long as the primary sequence is identical. The other atom pairs (*NL*2), which are of greater importance, will also sense this artificially amplified potential and will be biased toward 2.2 Å, a distance they rarely approach in reality (see Fig. S6).

In other words, both DFIRE and RW artificially tend to favor the extremely compact or compressed protein chains. This artifact is, however, greatly alleviated in SPOUSE, as suggested by the much shallower trough at 2.2 Å. Simultaneously, the two troughs located at ~7 and 8 Å are stabilized by a small amount in SPOUSE compared to DFIRE and RW. Therefore, SPOUSE generally favors the mildly compact protein chains but refuses to apply additional awards on the compressed chains. On the other hand, the SPOUSE backbone energy curve around 3–5 Å is elevated relative to DFIRE and RW. Elevation in this region will impair the stability of the local secondary structures (see Discussion). Fig. S5 *B* shows that backbone-sidechain interactions within 4–8 Å are stronger in SPOUSE than in DFIRE and RW, suggesting that SPOUSE favors denser backbone-sidechain packing. In terms of interactions between side-chain atoms, SPOUSE does not display much difference from DFIRE and RW (see Fig. S5, *C* and *D*).

Recognition of native conformation

Eight most widely used independent decoy sets are used to test the capability of SPOUSE to recognize the native conformations from decoys against DFIRE, DOPE, and RW. Table 1 lists the number of decoy sets in which the native conformation can be successfully selected based on the rule of lowest energy. For objective comparison, results

of decoy sets *4state_reduced* and *lmds* (shown in Table S2) are excluded from Table 1, because they have been used in the parameter optimization of SPOUSE. Except for the slightly worse result in *molder*, SPOUSE is the best in all decoy sets. The advantage is especially significant in *ROSETTA*, where SPOUSE succeeds in 27 out of the total 59 targets, at least five more than the other potentials. In total, SPOUSE correctly recognizes 114 native conformations out of the 152 decoy sets with a success rate of 75%. As a contrast, the success rates for DFIRE, DOPE, and RW are 71%, 68%, and 71%, respectively, lower than SPOUSE.

When a potential fails to recognize the native conformation by the rule of lowest energy, the rank of native conformation within the decoy set is another important factor to evaluate the power of this potential. A low rank means that the potential is challenged by very few misfolded conformers. Clearly, better potentials should be able to approach lower ranks in most decoy sets. The number in parentheses in Table 1 lists the average rank of native conformations in each decoy set when tested by different potentials. In this term, SPOUSE also wins in almost all decoy sets. After averaging over all decoy sets, SPOUSE has a mean rank of 10.7, better than DFIRE (13.1), DOPE (23.3), and RW (12.7). In summary, SPOUSE is significantly more powerful than the other major potentials in recognizing native conformations.

Notably, all potentials show indistinguishably good performance in the decoy sets *fisa*, *fisa_casp3*, *lattice_ssfit*, and *molder*, indicating the limited testing power of these sets. The evaluation results show largest variations in the decoy sets *ROSETTA* and *I-TASSER*, two sets generated from the most popular and successful protein structural prediction programs, ROSETTA and I-TASSER, respectively. Theoretically, decoy sets produced from real-time structural prediction programs are more valuable in evaluating the performance of statistical potentials. According to the recent study by Deng et al. (36), all statistical potentials are biased by decoy sets and no one is able to outperform the others in all decoy sets. Therefore, the testing behaviors in *ROSETTA* and *I-TASSER* should be considered more seriously, due to their higher relevance to practical protein structural prediction, also because *ROSETTA* and *I-TASSER* contains far more target proteins than the other decoy sets. Notably, SPOUSE has the highest rank in both decoy sets.

Selection of the close-to-native models

Although recognition competence of native structures is a quite important standard in potential evaluation, it is somehow of limited usefulness in practice. The absence of native structure during a blind structure prediction process requires that a good potential should be able to select close-to-native conformers solely by the criterion of energy (9,10,19). These selected conformers could be further refined to improve their similarity to the native structure by the structural refinement programs. As proposed by previous groups, the real-time simulation and structure prediction-based *ROSETTA* and *I-TASSER* decoy sets are more realistic and challenging (19). Thus, we use them as the test sets to evaluate the selection of close-to-native models in this work.

The energetically most favorable top 1, 5, and 10 conformers are extracted from each decoy sets and the α-carbon root-mean-square-distance (CαRMSD) of the best model relative to the native conformation is listed in Table 2. A good potential should approach small CαRMSD values in all top 1/5/10 tests. In *ROSETTA* decoy sets, SPOUSE performs best in top5 and top10 tests and ranks third in top1 test. In *I-TASSER* decoy sets, SPOUSE wins top1 and top5 test and ranks third in top10 test. After naively averaging the performance of potential among the six tests, SPOUSE has a mean rank of 1.67, lower than the values of DFIRE, DOPE, and RW, which are 2.50, 3.17, and 2.33, respectively.

Despite the good relative performance of SPOUSE shown above, the mean CαRMSD values obtained from different potentials are quite close, suggesting the limited discriminating power of this test. This can be further demonstrated by the overlap of error bars in the statistical analysis conducted in a bootstrap strategy (see Table S3), where the test was repeated in 10 decoy subsets obtained by randomly removing 20% structures from the corresponding decoy set.

Correlation coefficient between energy and model errors

In addition to the ability to select close-to-native conformer, an ideal potential should possess the following property: the conformations with lower energy should be closer to the native structure, or have lower RMSDs (or higher TM-scores). Thus, the overall energy landscape will be inclined to the native conformation such that it naturally guides the chain to fold into the native structure in molecular simulation. This inclination is most frequently quantitatively described by the Pearson correlation coefficient (CC) between energy and the relative distance (RMSD or TM-score) to the native conformation. Decoy sets of good and poor CC are shown in Fig. S7 as examples.

The correlation coefficients of DFIRE, DOPE, RW, and SPOUSE are calculated for the *ROSETTA* and *I-TASSER*

TABLE 2 Selection of close-to-native models

Best models		DFIRE	DOPE	RW	SPOUSE
ROSETTA	Top1	7.33	**7.16**	7.46	7.36
	Top5	5.53	5.58	**5.50**	**5.50**
	Top10	5.27	5.28	5.24	**5.20**
I-TASSER	Top1	5.25	5.29	5.20	**5.16**
	Top5	4.38	4.33	4.36	**4.25**
	Top10	**3.79**	4.01	3.80	3.94

Top1, top5, or top10 are the top ranking 1, 5, or 10 models with lowest calculated energy by the corresponding potentials, respectively. Numbers are the smallest CαRMSD in the top models. The best performance in each category is highlighted in bold.

decoy sets, and are then listed in Table 3. SPOUSE shows highest CC value (0.460 or −0.472) in *ROSETTA*. According to the statistical test conducted in a bootstrap strategy (see Table S4), SPOUSE's prominent performance is statistically significant. In addition, both DFIRE and RW approach higher CC values (~0.45 or −0.46) in this work than those listed in the literature (~0.44 or −0.43), because of the rapid increase of available crystal structures in PDB. In *I-TASSER*, SPOUSE is worse than DFIRE and RW, with comparable performance to DOPE (Table 3 and see Table S4). The reason of its relative weak performance will be analyzed in the Discussion section.

DISCUSSION

About the reference state

To strictly analyze the influence of the reference state on the performance of the statistical potential, SPOUSE, DFIRE, and RW potentials are calculated using the same crystal structure database, the same cutoff distance, and the same repulsive penalty at short distance. In other words, all factors other than the reference state are forced identical to exclude their effects. Therefore, the advantage of SPOUSE over DFIRE and RW in most tests can be ascribed to its better modeling of the reference state. As described in Results, the distinction between SPOUSE and DFIRE/RW is most conspicuous at short distance (<4 Å in Fig. 1), where the SPOUSE reference curve is elevated significantly, which weakens short-range backbone-backbone interaction and makes the energy well shallower (see Fig. S5 *A*). We believe that this elevation results from the correct weighting of the local and nonlocal distance distribution (see Eq. 7).

Let us suppose two residues separated by k residues in an N-residue chain, and denote the distribution of the distance between their α-carbons as $P(r|k)$. Then distribution of distance between any α-carbon pairs can be obtained by simply averaging $P(r|k)$ of all available k values ($k = 0, 1, \ldots, N\text{-}2$), by assuming that residue pairs are equally probable to be separated by k residues. This is actually the procedure RW took in deriving the expected probability. However, the above assumption is only valid when the chain is infinitely long. In fact, in a finite N-residue chain, the residue pairs separated by k residues appear (N-k-1) times altogether. After dividing by the overall number of residue pairs $N(N\text{-}1)/2$, the probability of observing a k-residue-separated pair is

$$P(k) = \frac{N-k-1}{N(N-1)/2},$$

which diminishes linearly with the increase of k. Therefore, the distribution of local pairs should be assigned more weights than the nonlocal pairs (as processed by SPOUSE in Eq.7), and this leads to the elevation of SPOUSE reference curve at short distance. To verify this idea, we rederived the expected probability of SPOUSE (see Eq. 7) by assuming the equal probability of all CA-CA pairs (similar to RW procedure), and plotted the curve (*triangle-curve*) in Fig. 1. The curve falls toward the RW reference curve significantly at short distance.

Correction on the weighting factors, however, brings an unexpected side effect: the potential minima between 3 and 5 Å in SPOUSE backbone potential is also elevated, which leads to the destabilization of local secondary structures and the drop of performance in decoy sets where backbone secondary structures are poorly assembled. The relatively low performance of SPOUSE in the correlation coefficient test in *I-TASSER* decoy set is caused by this reason. For verification, the losses of secondary structures in decoys relative to the native structures are estimated in both *ROSETTA* and *I-TASSER* decoy sets, respectively, by

$$\frac{SS_{native} - SS_{decoy}}{SS_{native}} \times 100\%,$$

where SS is the total number of noncoil residues according to the secondary structures assigned by DSSP (36). As shown in Fig. S8, *I-TASSER* decoys on average lose 40% secondary structures. As a contrast, the secondary structure contents are retained in the *ROSETTA* decoy set, possibly because hydrogen-bonding potential is implemented in the ROSETTA algorithm to optimize the local backbone structures. Therefore, the negative effect of SPOUSE on secondary structure assembly can be rescued by explicit hydrogen bonds (see the high CC values of SPOUSE in the *ROSETTA* decoy sets).

TABLE 3 Correlation coefficients in the *ROSETTA* and *I-TASSER* decoy sets

Decoy set	DFIRE	DOPE	RW	SPOUSE
ROSETTA	0.449/−0.457	0.431/−0.424	0.451/−0.463	**0.460/−0.472**
I-TASSER	**0.517**/−0.497	0.479/−0.487	0.515/−**0.507**	0.497/−0.480

The values before and after the slash are the ones calculated from RMSD and TM-score, respectively. The best performance in each category is highlighted in bold.

Influence of the cutoff distance

Theoretically, greater cutoff distance results in better performance by including more information. However, this is not the case for most distance-dependent statistical potentials. Many studies focused on the influence of cutoff distance (13,14,50), but no views have been commonly accepted yet. In practice, most statistical potentials adopt a 15 Å cutoff. In this work, we compared the performance of SPOUSE with those of DFIRE and RW at various cutoff distances (from 14 Å to 20 Å) on all decoy sets.

The influence of cutoff distances on potentials is illustrated in Fig. 2. In all potentials, a larger cutoff distance results in poorer performance in identifying native

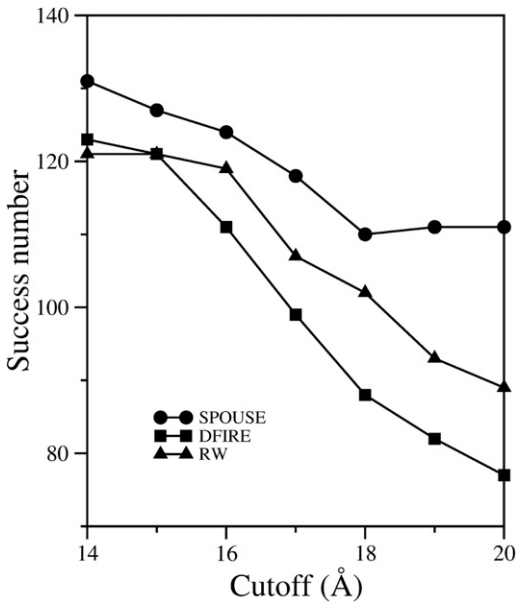

FIGURE 2 Influence of the cutoff distance on different potentials. The success number of SPOUSE (*circle line*) is higher than DFIRE (*square line*) and RW (*triangle line*) at the whole range tested. Furthermore, it decreases, at a much slower pace than DFIRE and RW, illustrating a more robust performance at the variation of cutoff distance.

conformations. However, this weakening of performance is most serious in DFIRE, because its reference state curve increases rapidly at long distance (beyond 15 Å) and violates the expectation that the distance between two atoms in a polymer chain cannot exceed the end-to-end distance. The performance of SPOUSE declines at a much slower pace at large cutoff distances (~20 Å), because it better models the reference state using a polypeptide chain. As expected, RW performs at an intermediate level between SPOUSE and DFIRE, because partial information about the connectivity is included in its reference state. In summary, SPOUSE is more robust upon the change of cutoff distance than the other distance-dependent statistical potentials.

Possibility of improvement by orientation-dependent side chains

Because composite statistical potentials are always better than single-attribute-dependent potentials (51), many distance-dependent potentials developed their composite counterparts. Among them, GOAP (20) is a successful example by combining the DFIRE potential and a self-developed orientation-dependent side-chain potential named as GOAP-AG. The orientation-dependent side-chain potentials usually can greatly improve the overall performance by including the many-body effects rather than only considering pairwise interactions (19,52). To speculate the future possibility of SPOUSE to integrate with an orientation-dependent side-chain potential, we combined SPOUSE and GOAP-AG (SPOUSE +GOAP-AG) and tested its performance against GOAP (or DFIRE+GOAP-AG) on all the eight independent decoy sets mentioned above except *lattice_ssfit* (in which only a minor part of the decoy models are calculable by the downloaded GOAP program).

TABLE 4 Performance of orientation-dependent SPOUSE and DFIRE

Decoys	DFIRE+ GOAP-AG	SPOUSE+ GOAP-AG	SPOUSE	Targets
4state_reduced	7/0.732	7/0.734	6/0.551	7
lmds	7/0.123	7/0.134	7/0.126	10
fisa	3/0.268	3/0.302	3/0.283	4
fisa_casp3	5/0.167	5/0.175	5/0.261	5
molder	19/0.823	19/0.841	18/0.769	20
ROSETTA	47/0.488	48/0.495	27/0.460	59
I-TASSER	45/0.449	47/0.464	53/0.497	56
Total/average	133/0.489	136/0.500	119/0.484	161

Orientation-dependence integrated SPOUSE (SPOUSE + GOAP-AG) and the original GOAP (DFIRE + GOAP-AG) are compared. The numbers before the slash are the numbers of selected native conformations, and the ones after the slash are the corresponding Pearson correlation coefficients. The corresponding number for SPOUSE alone was appended at right to facilitate comparison.

Table 4 compares the performance of SPOUSE+GOAP-AG versus GOAP (or DFIRE+GOAP-AG) in the test of native structure recognition (former) and the test of correlation coefficient (latter). The data for SPOUSE alone is also listed for easy comparison. As a whole, after the integration of GOAP-AG, SPOUSE is significantly improved. In particular, it correctly selected 48 native conformations out of 59 *ROSETTA* targets, 21 more than before (SPOUSE alone). At the same time, the correlation coefficient also rises greatly (from 0.460 to 0.495). More importantly, SPOUSE+GOAP-AG potential prevails GOAP in all decoy sets tested, in both the number of successful native recognition and the correlation coefficients. This convinces us that SPOUSE could be sufficiently improved by the integration of a suitable orientation-dependent side-chain potential.

SUPPORTING MATERIAL

Four tables and eight figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)01060-0.

The authors gratefully thank Xinqi Gong for discussion and suggestions and James Hinshaw for his kind help in generating the unfolded state ensembles.

This work was supported by the Tsinghua-Yue-Yuen Medical Sciences Fund and Tsinghua National Laboratory for Information Science and Technology.

REFERENCES

1. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein DataBank. *Nucleic Acids Res.* 28:235–242.

2. Apweiler, R., A. Bairoch, …, L. S. Yeh. 2004. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32(Database issue): D115–D119.
3. Zhang, Y. 2009. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.* 19:145–155.
4. Fitzgerald, J. E., A. K. Jha, …, K. F. Freed. 2007. Reduced C(β) statistical potentials can outperform all-atom potentials in decoy identification. *Protein Sci.* 16:2123–2139.
5. Brooks, B. R., C. L. Brooks, 3rd, …, M. Karplus. 2009. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.
6. Case, D. A., T. E. Cheatham, 3rd, …, R. J. Woods. 2005. The AMBER biomolecular simulation programs. *J. Comput. Chem.* 26:1668–1688.
7. Christen, M., P. H. Hünenberger, …, W. F. van Gunsteren. 2005. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* 26:1719–1751.
8. Jorgensen, W. L., and J. Tirado-Rives. 1988. The OPLS potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110:1657–1671.
9. Shen, M. Y., and A. Sali. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507–2524.
10. Rykunov, D., and A. Fiser. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics.* 11:128.
11. 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.
12. Sippl, M. J. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* 7:473–501.
13. Samudrala, R., and J. Moult. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* 275:895–916.
14. Lu, H., and J. Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins.* 44:223–232.
15. Zhou, H., and Y. Zhou. 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.
16. Zhang, C., S. Liu, …, Y. Zhou. 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* 13:400–411.
17. Ferrada, E., I. A. Vergara, and F. Melo. 2007. A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. *Cell Biochem. Biophys.* 49:111–124.
18. Mirzaie, M., C. Eslahchi, …, M. Sadeghi. 2009. A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys. *Proteins.* 77:454–463.
19. Zhang, J., and Y. Zhang. 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.
20. Zhou, H., and J. Skolnick. 2011. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101:2043–2052.
21. Tanaka, S., and H. A. Scheraga. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules.* 9:945–950.
22. Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.
23. McConkey, B. J., V. Sobolev, and M. Edelman. 2003. Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. USA.* 100:3215–3220.
24. Arab, S., M. Sadeghi, …, A. Sheari. 2010. A pairwise residue contact area-based mean force potential for discrimination of native protein structure. *BMC Bioinformatics.* 11:16.
25. Ellison, P. A., and S. Cavagnero. 2006. Role of unfolded state heterogeneity and en-route ruggedness in protein folding kinetics. *Protein Sci.* 15:564–582.
26. Fleming, P. J., and G. D. Rose, editors. 2005. Conformational Properties of Unfolded Proteins. Wiley-VCH, Weinheim, Germany. 710–736.
27. Schweitzer-Stenner, R. 2012. Conformational propensities and residual structures in unfolded peptides and proteins. *Mol. Biosyst.* 8:122–133.
28. Millett, I. S., S. Doniach, and K. W. Plaxco. 2002. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Adv. Protein Chem.* 62:241–262.
29. Kohn, J. E., I. S. Millett, …, K. W. Plaxco. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12491–12496.
30. Shortle, D., and M. S. Ackerman. 2001. Persistence of native-like topology in a denatured protein in 8 M urea. *Science.* 293:487–489.
31. Möglich, A., K. Joder, and T. Kiefhaber. 2006. End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci. USA.* 103:12394–12399.
32. Fitzkee, N. C., and G. D. Rose. 2004. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12497–12502.
33. Pappu, R. V., R. Srinivasan, and G. D. Rose. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA.* 97:12565–12570.
34. Shi, Z., K. Chen, …, N. R. Kallenbach. 2006. Conformation of the backbone in unfolded proteins. *Chem. Rev.* 106:1877–1897.
35. Jha, A. K., A. Colubri, …, T. R. Sosnick. 2005. Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA.* 102:13099–13104.
36. Deng, H., Y. Jia, …, Y. Zhang. 2012. What is the best reference state for designing statistical atomic potentials in protein structure prediction? *Proteins.* 80:2311–2322.
37. Yang, Y., and Y. Zhou. 2008. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.* 17:1212–1219.
38. Park, B., and M. Levitt. 1996. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392.
39. Keasar, C., and M. Levitt. 2003. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* 329:159–174.
40. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 77:778–795.
41. Wang, G., and R. L. Dunbrack, Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics.* 19:1589–1591.
42. Simons, K. T., C. Kooperberg, …, D. Baker. 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.
43. Samudrala, R., Y. Xia, …, E. S. Huang. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.* 4:505–516.
44. Xia, Y., E. S. Huang, …, R. Samudrala. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* 300:171–185.
45. Samudrala, R., and M. Levitt. 2000. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.* 9:1399–1401.

46. Eramian, D., M. Y. Shen, …, M. A. Marti-Renom. 2006. A composite score for predicting errors in protein structure models. *Protein Sci.* 15:1653–1666.

47. Qian, B., S. Raman, …, D. Baker. 2007. High-resolution structure prediction and the crystallographic phase problem. *Nature*. 450: 259–264.

48. Eswar, N., B. Webb, …, A. Sali. 2006. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*. Chapter 5:Unit 5.6.

49. Reference deleted in proof.

50. Melo, F., R. Sánchez, and A. Sali. 2002. Statistical potentials for fold assessment. *Protein Sci.* 11:430–448.

51. Melo, F., and A. Sali. 2007. Fold assessment for comparative protein structure modeling. *Protein Sci.* 16:2412–2426.

52. Yang, Y., and Y. Zhou. 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins.* 72:793–803.