Analyses on hydrophobicity and attractiveness of all-atom distance-dependent potentials

ARTICLE *in* PROTEIN SCIENCE · SEPTEMBER 2009

Impact Factor: 2.85 · DOI: 10.1002/pro.201 · Source: PubMed

3 AUTHORS, INCLUDING:

Takashi Ishida
Tokyo Institute of Technology
35 PUBLICATIONS **598** CITATIONS
SEE PROFILE

Kengo Kinoshita
94 PUBLICATIONS **2,659** CITATIONS
SEE PROFILE

Analyses on hydrophobicity and attractiveness of all-atom distance-dependent potentials

Matsuyuki Shirota,[1,2] Takashi Ishida,[1] and Kengo Kinoshita[1,2]*

[1]Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

[2]Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

Received 28 April 2009; Revised 17 June 2009; Accepted 18 June 2009
DOI: 10.1002/pro.201
Published online 8 July 2009 proteinscience.org

Abstract: Accurate model evaluation is a crucial step in protein structure prediction. For this purpose, statistical potentials, which evaluate a model structure based on the observed atomic distance frequencies in comparison with those in *reference states*, have been widely used. The reference state is a virtual state where all of the atomic interactions are turned off, and it provides a standard to measure the observed frequencies. In this study, we examined seven all-atom distance-dependent potentials with different reference states. As results, we observed that the variations of atom pair composition and those of distance distributions in the reference states produced systematic changes in the hydrophobic and attractive characteristics of the potentials. The performance evaluations with the CASP7 structures indicated that the preference of hydrophobic interactions improved the correlation between the energy and the GDT-TS score, but decreased the *Z*-score of the native structure. The attractiveness of potential improved both the correlation and *Z*-score for template-based modeling targets, but the benefit was smaller in free modeling targets. These results indicated that the performances of the potentials were more strongly influenced by their characteristics than by the accuracy of the definitions of the reference states.

Keywords: structure-sequence relationship; statistical potential; model evaluation; hydrophobicity; attractive interaction; CASP

Introduction

The prediction of a protein structure from its sequence is still a very challenging task in structural bioinformatics. Because of the rapid growth in computational power, we are now able to enumerate a large number of model structures, but it is still difficult to select a few suitable models from them.[1] Therefore, it is a fundamental problem to establish an evaluation function to measure the fitness of model structures for reliable structure prediction. As analyses of known protein structure databases will provide abundant information about the structure-sequence relationships, the known protein structures in the Protein Data Bank[2] have been used to derive evaluation functions, which are called statistical potentials.[3–6]

In the construction of statistical potentials, three major factors should be determined: structural representation (e.g., amino acid residues or all atoms), structural variables (e.g., solvent accessibilities, secondary structures, and pair distances), and a reference state. For the structural variables, distance-dependent

Abbreviations: AKBP, atomic knowledge-based potential; CASP, critical assessment of techniques for protein structure prediction; DFIRE, distance-scaled finite ideal-gas reference state; RAPDF, residue-dependent atomic probability density function.

Grant sponsor: Grant-in-Aid for Scientific Research on the Priority Area Transportsome from Ministry of Education, Culture, Sports, Science, and Technology of Japan; Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation; Annual Budget of Human Genome Center.

*Correspondence to: Kengo Kinoshita, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
E-mail: kino@ims.u-tokyo.ac.jp

all-atom representation is currently considered to be more accurate than those with residue representation and/or distance-independent variables.[7–9] Actually, all-atom distance-dependent potentials are frequently used for protein structure predictions[7,8] along with a few other applications, such as prediction of mutation-induced stability changes,[9] loop modeling,[10] and assessment of protein–protein interactions.[11] Therefore, we focused on this type of potential in this study.

A reference state is a virtual state, where all atomic interactions are turned off. More formally, suppose that the interaction energy of an atom pair ab at a distance r, $E_{ab}(r)$, is calculated by the following equation,[12]

$$E_{ab}(r) = -kT \ln \frac{N_{ab}^{\text{observed}}(r)}{N_{ab}^{\text{expected}}(r)}, \quad (1)$$

where k, T, and $N_{ab}^{\text{observed}}(r)$ are the Boltzmann constant, the absolute temperature, and the number of observed atom pairs ab at a distance r in the set of native structures, respectively. The value $N_{ab}^{\text{expected}}(r)$ is regarded as a reference state and provides a basis to measure the relative abundance of $N_{ab}^{\text{observed}}(r)$. Although $N_{ab}^{\text{observed}}(r)$ can be obtained simply by monitoring the occurrence, their expected values or the "observation in the reference state," $N_{ab}^{\text{expected}}(r)$, may vary with the assumption used in each study, because the reference state is a virtual state, and the exact evaluation of its possible frequency is impossible. As a consequence, several different types of reference states have been proposed.[7,10,13–16]

When these statistical potentials were developed, the analyses were performed by focusing on the theoretical accuracy of the reference states and the performance of the potentials in the evaluation of model structures. However, the relationship among the different reference states has been rarely discussed, as compared with the analyses in residue-level potentials, where the preference for the clustering of hydrophobic residues in the protein core has been shown to play a dominant role in their performance.[17] In addition, the characterization of the various reference states that have been proposed thus far was not performed well, except for a few remarkable pioneering efforts. For example, Zhou and Zhou[14] reported that their reference states have attractive characteristics; that is, all of the atom pairs in the middle range distance will have favorable energies regardless of their atom types, and they argued that the attractiveness could improve the performance of their potential.

In this study, we considered seven possible reference states, four developed by other groups[14–16,18] and three new ones, and we clarified the effects of each reference state on the structure prediction performance by using CASP7 models. As a result, both the attractive and hydrophobic characteristics of the potentials were introduced by some types of reference states, and they systematically changed their performance.

Results and Discussion

The formulation of the seven reference states

Because the exact evaluation of the atom pair frequencies in the reference state is not possible, they have been calculated based on some assumptions described below. In the native state, each atom pair has its specific distance dependence according to their atom types, which is characteristic of the native protein conformation. On the other hand, if the atomic interactions were turned off in the reference states, all of the atom pairs will have the same distance dependence. This leads to the same composition of atom pairs at any distance in the reference state. Thus, the expected number of atom pairs ab at a distance r in the reference state can be simply defined as,

$$N_{ab}^{\text{expected}}(r) = N \times p^{\text{expected}}(r) \times p^{\text{expected}}(ab), \quad (2)$$

where N is the total number of atom pairs in the data set of protein domains, $p^{\text{expected}}(r)$ is the probability of atom pairs at a distance r, and $p^{\text{expected}}(ab)$ is the probability of atom pairs ab in the reference state. In this study, several definitions of $p^{\text{expected}}(r)$ and $p^{\text{expected}}(ab)$ were tested in combinations, resulting in the seven reference states.

The first reference state was RAPDF (residue-dependent atomic probability density function), which was introduced by Samudrala and Moult.[16] In this reference state, $p^{\text{expected}}(r) \equiv p^{\text{observed}}(r) = N^{\text{observed}}(r)/N$, and

$$p^{\text{expected}}(ab) \equiv p^{\text{observed}}(ab|r \le r_{\text{cut}}) = \sum_{r}^{r_{\text{cut}}} N_{ab}^{\text{observed}}(r) \Big/ \sum_{ab} \sum_{r}^{r_{\text{cut}}} N_{ab}^{\text{observed}}(r),$$

where $N^{\text{observed}}(r) = \sum_{ab} N_{ab}^{\text{observed}}(r)$. The cut-off distance r_{cut} is the upper limit of the distance within which the atom pair interaction is considered.

The second reference state was DFIRE (distance-scaled finite ideal-gas reference state), which was developed by Zhou and Zhou.[14] They considered that the atom pairs at the cut-off distance were noninteracting pairs.

$$p^{\text{expected}}(ab) \equiv p^{\text{observed}}(ab|r = r_{\text{cut}}) = N_{ab}^{\text{observed}}(r_{\text{cut}}) \Big/ \sum_{ab} N_{ab}^{\text{observed}}(r_{\text{cut}}).$$

In this reference state,

$$p^{\text{expected}}(r) \equiv p^{\text{observed}}(r_{\text{cut}}) \times 4\pi r^{\alpha} \Delta r / 4\pi r_{\text{cut}}^{\alpha} \Delta r_{\text{cut}}.$$

They argued that the number of occurrences of atom pairs at a distance r in the reference state should increase with r^2 if the proteins were infinitely large;

Table I. *The Seven Reference States and Their Characteristics*

	Definition		Characteristics	
Potential	$p^{\text{expected}}(r)$	$p^{\text{expected}}(ab)$	Attractive	Hydrophobic
RAPDF	$p^{\text{observed}}(r)$	$p^{\text{observed}}(ab\|r \le r_{\text{cut}})$	−	−
DFIRE	$p^{\text{observed}}(r_{\text{cut}}) \times \frac{4\pi r^{\alpha}\Delta r}{4\pi r_{\text{cut}}^{\alpha}\Delta r_{\text{cut}}}$	$p^{\text{observed}}(ab\|r = r_{\text{cut}})$	++	+
SHELL	$p^{\text{observed}}(r)$	$p^{\text{observed}}(ab\|r = r_{\text{cut}})$	−	+
BALL	$p^{\text{observed}}(r \le r_{\text{cut}}) \times \frac{4\pi r^{\alpha}\Delta r}{\sum_{r}^{r_{\text{cut}}} 4\pi r^{\alpha}\Delta r}$	$p^{\text{observed}}(ab\|r \le r_{\text{cut}})$	+	−
AKBP	$p^{\text{observed}}(r)$	$\chi_a\chi_b$	−	++
FAR	$p^{\text{observed}}(r)$	$p^{\text{observed}}(ab\|r > r_{\text{cut}})$	−	+++
DS	$p^{\text{observed}}(r)$	$\frac{N_{ax}(r)\times N_{bx}(r)}{N(r)^2}$	−	−

The first column shows the potential names. The second and third columns are the definitions of the reference states on the distance dependence ($p^{\text{expected}}(r)$) and on the atom pair composition ($p^{\text{expected}}(ab)$), respectively. The fourth and fifth columns are the attractive and hydrophobic characteristics of the potentials, respectively, where + indicates that a potential has the corresponding characteristics, and − means it does not. The number of "+" signs roughly indicates the strength of the characteristic. The definitions of distance dependence in the second column influence the attractive characteristics in the fourth column, whereas those of the atom pair composition in the third column influence the hydrophobic characteristics in the fifth column (see the text for details).

however, because the size of the protein molecule was finite, it should increase with r^{α}, with α smaller than 2. In this study, we used the value of α, 1.61, defined by Zhou and Zhou.

Because RAPDF and DFIRE differ with each other in both $p^{\text{expected}}(r)$ and $p^{\text{expected}}(ab)$, we constructed two new reference states, as the hybrid reference states. The third reference state used the distance dependence in RAPDF, $p^{\text{expected}}(r) \equiv p^{\text{observed}}(r)$, and the atom pair composition in DFIRE, $p^{\text{expected}}(ab) \equiv p^{\text{observed}}(ab|r = r_{\text{cut}})$. We call this potential SHELL, because it uses the atom pair composition at the shell of the cut-off distance.

The fourth potential uses the distance dependence in DFIRE,

$$p^{\text{expected}}(r) \equiv \left(\sum_{r=0}^{r_{\text{cut}}} p^{\text{observed}}(r)\right) \times 4\pi r^{\alpha}\Delta r \Big/ \sum_{r=0}^{r_{\text{cut}}} 4\pi r^{\alpha}\Delta r,$$

and the atom pair composition in RAPDF, $p^{\text{expected}}(ab) \equiv p^{\text{observed}}(ab|r \le r_{\text{cut}})$. Here, the same value of α, 1.61, as in DFIRE was used. We call this potential BALL, because it uses the atom pairs at a distance r_{cut} or less.

For the three remaining reference states, $p^{\text{expected}}(r) \equiv p^{\text{observed}}(r)$, but $p^{\text{expected}}(ab)$ varied with each other. The fifth reference state was AKBP (atomic knowledge-based potential), proposed by Lu and Skolnick,[15] in which $p^{\text{expected}}(ab) \equiv \chi_a\chi_b$. χ_a indicated the mole fraction of atom type a. This atom pair composition would be close to that of all of the atom pairs in the database irrespective of the distance between them; that is,

$$\chi_a\chi_b \cong \sum_{r=0}^{\infty} N_{ab}^{\text{observed}}(r) \Big/ \sum_{r=0}^{\infty}\sum_{ab} N_{ab}^{\text{observed}}(r).$$

We defined another reference state that uses the atom pair composition at a distance larger than the cutoff, $p^{\text{expected}}(ab) \equiv p^{\text{observed}}(ab|r > r_{\text{cut}})$, because distant atom pairs would be noninteracting. This potential was called FAR in this study.

The seventh reference state, called DS, was proposed by DeBolt and Skolnick.[18] It assumes that the occurrence of atom pair ab at a distance r should be proportional to the product of the number of atom pairs between a and any other atom and the number of atom pairs between b and any other atom, that is, $p^{\text{expected}}(ab) \equiv N_{ax}^{\text{observed}}(r) \times N_{bx}^{\text{observed}}(r)/N^{\text{observed}}(r)^2$, where $N_{\text{ax}}^{\text{observed}}(r)$ is the number of atom pairs between a and any atom at a distance r.

These definitions of the seven reference states are summarized in Table I.

Distance dependence of $p^{\text{expected}}(r)$

We first considered the distance dependence of the reference states, where three possible formulations are discussed, that is (a) $p^{\text{expected}}(r) \equiv p^{\text{observed}}(r)$, (b) $p^{\text{observed}}(r_{\text{cut}}) \times 4\pi r^{\alpha}\Delta r/4\pi r_{\text{cut}}^{\alpha}\Delta r_{\text{cut}}$, and (c) $p^{\text{observed}}(r \le r_{\text{cut}}) \times 4\pi r^{\alpha}\Delta r/\sum_{r}^{r_{\text{cut}}} 4\pi r^{\alpha}\Delta r$.

The first definition (a) was the most popular one and was used in the potentials RAPDF, SHELL, AKBP, FAR, and DS. Definition (a) originally came from the "uniform density" reference state, proposed by Sippl[12] for the residue-level distance-dependent potential. On the other hand, in the second and the third definitions ((b) and (c)), which were used for DFIRE and BALL, respectively, the expected probabilities were not always identical to the observed probabilities. This biased reference state was proposed by Zhou and Zhou,[14] based on the argument that the occurrence of atom pairs at a distance r in the reference state should increase with r^{α} in real proteins of finite sizes, where α is smaller

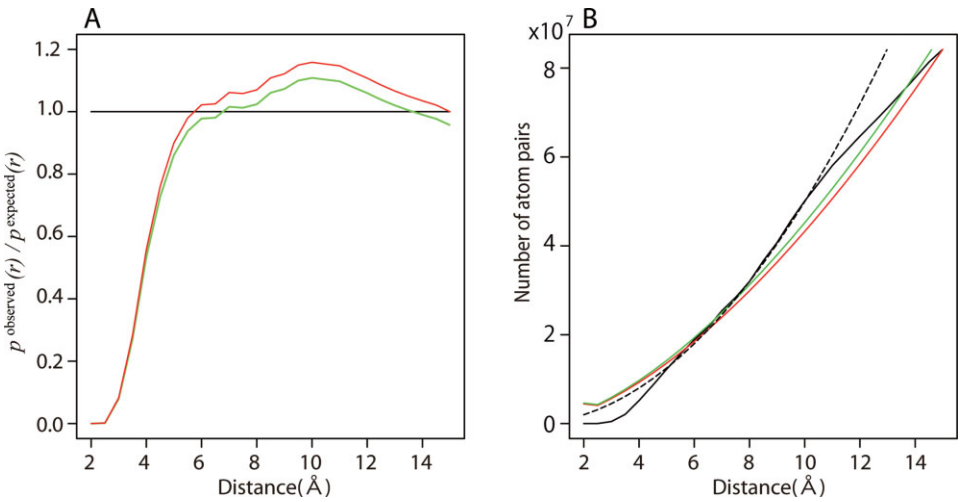

Figure 1. (A) The ratio between the observed probabilities of the atom pairs at each distance and the expected ones for DFIRE (red), BALL (green), and the other five potentials (black). (B) The observed atom pair occurrence at each distance in the dataset (black solid line) and the expected ones for DFIRE (red) and BALL (green). The dotted line is $N = ar^2$, which crosses the black line (observations) at a distance of 8 Å. The number of observed atom pairs increases approximately proportional to r^2 at distances <10 Å, but it remarkably deviates from the dotted line at larger distances.

than 2, from the analogy that the number of atom pairs increases with r^2 in an ideal gas in infinite space. They determined the value of α to be 1.61, so that it minimized the relative fluctuation of the function $N(r)/r^{\alpha}$ within the range of 0–15 Å, where $N(r)$ was the number of atom pairs at the distance r in the uniformly distributed points in a finite sphere. In this study, we used the same value of α for the potentials DFIRE and BALL.

To illustrate the features of each reference state, we first observed the ratio $p^{\text{observed}}(r)$ to $p^{\text{expected}}(r)$ [Fig. 1(A)]. The red and green lines are the plots for DFIRE and BALL, respectively, and the horizontal black line is for the other five potentials. As the negative log of this ratio corresponds to the potential energy [Eq. (2)], the ratio >1.0 indicates that the potential energy yields negative values, or that the interactions in the range with the ratio >1.0 are considered to be favorable. The observed probabilities were larger than the expected probabilities in the range from 6 to 14 Å for DFIRE (red line). This was essentially the same result as that shown in Figure 5

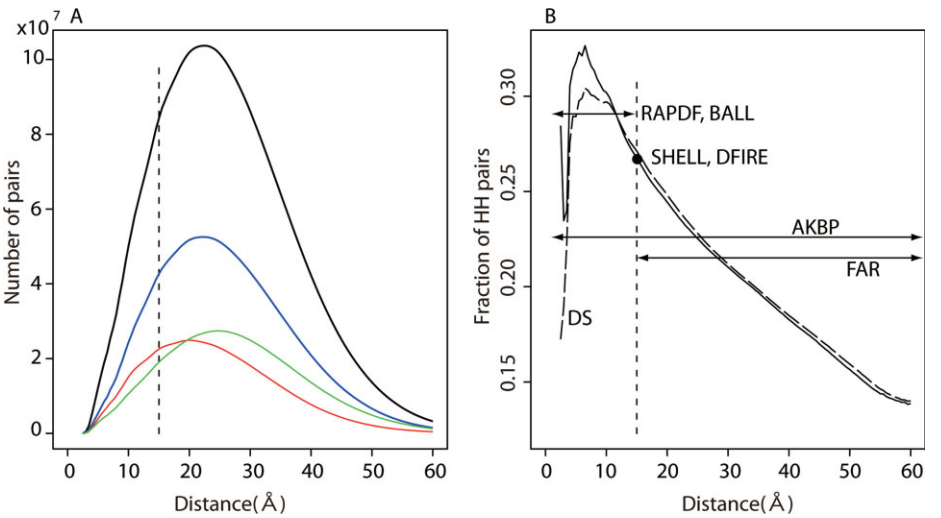

Figure 2. Distribution of atom pairs as a function of distance. (A) The numbers of atom pairs in the database as functions of distance for all atom types (black line), hydrophobic–hydrophobic pairs (HH pairs, red), polar–polar pairs (green), and hydrophobic–polar pairs (blue), where ACGILMFWYV were treated as hydrophobic residues and all other residues were considered as polar residues. The dotted vertical line is drawn at the distance = 15 Å, which was the cut-off distance used in this study to determine the interacting atom pairs. (B) The fraction of atom pairs from HH residue pairs. The solid line represents the fraction of HH atom pairs at each distance. The fractions of HH pairs in each potential other than SHELL, DFIRE, and DS are shown in horizontal lines in the range of each reference state. The filled circle at 15 Å represents the fractions of HH pairs in SHELL and DFIRE, and the dashed line shows the fraction of HH pairs of DS.

in Zhou and Zhou.[14] In a similar way, the potential BALL (green line) will evaluate the atom pairs within the distance of 7 Å $< r <$ 13 Å favorably. Since the distance range of BALL covered a major part (about 77%) of all of the atom pairs, BALL and DFIRE possess a kind of bias, in that the atom pairs in this range are considered favorable regardless of the atom types. We call this feature the attractiveness of a potential. For the other five potentials, at any distance the attractive and repulsive interactions cancel each other over all of the atom pairs (black line), and thus there is no attractive bias. The attractiveness of the seven potentials is summarized in Table I.

The attractiveness of DFIRE and BALL results from the fact that they assumed that the probability of the atom pairs increased with a constant exponent of α as a function of the distance r, but $p^{\text{observed}}(r)$ actually increases with a larger exponent than α for small r, and the increase becomes slower for large r [Fig. 1(B)]. The introduction of the attractiveness of the potential can be beneficial in evaluating model structures, because native structures generally have compact globular structures. It should be noted, however, that it has no theoretical reasons and it is based on practical consequences.

Atom pair composition in the reference states

Another major difference to be considered is the atom pair composition of reference states, and several ideas have been proposed, as formally described earlier and in Table I. While RAPDF and BALL use the atom pair frequencies within a distance $r \leq r_{\text{cut}}$, DFIRE and SHELL use that just at the distance $r = r_{\text{cut}}$ and FAR considers that with a distance $r > r_{\text{cut}}$. AKBP uses the atom pair composition regardless of the distance, while DS tries to estimate the observed distributions at each distance. The hydrophobic residues tend to cluster inside the molecule, and thus it is expected that the atom pairs between hydrophobic residues (HH pairs) will occupy larger fractions in a shorter distance, which will change the atom pair composition of each reference state.

Figure 2(A) shows how the distributions of the hydrophobic pairs and hydrophilic pairs differ by distance in our dataset. The black line is the number of all atom pairs at each distance, while the red, blue, and green lines represent the atom pairs between HH pairs, HP (hydrophobic–hydrophilic) pairs, and PP (hydrophilic–hydrophilic) pairs, respectively. The distribution of HH pairs is clearly shifted toward the shorter distance, as compared with those of the HP or PP pairs. This fact will produce the difference in the hydrophobic characteristics between potentials. In Figure 2(B), the fraction that HH pairs occupy at each distance is plotted as a black solid line, together with the distance range and the fraction of HH pairs in the reference state for each potential, represented by the horizontal lines, the filled circle, or the dashed line. It should be noted that the atom pairs within the distance of 15 Å, which are the target of the energy evaluation in this study, are highly biased toward HH pairs. The lower the fraction of HH pairs in the reference state, the lower their energies will be. We call this tendency of the potential to evaluate HH pairs favorably as the hydrophobic characteristic, and the seven potentials can be arranged in the ascending order of hydrophobic characteristics as RAPDF, BALL $<$ DS $<$ SHELL, DFIRE $<$ AKBP $<$ FAR [Fig. 2(B), an average value $<$15 Å was used for DS]. Considering the strong contribution of the hydrophobic effect in the model evaluation, the difference between these potentials would not be small. Because of the lack of the correct observation in the reference state, we cannot decide which one is superior to another, or what the correct hydrophobicity of the reference state is, in theoretical terms.

Effect of the reference states on the atom pair interaction energies

To illustrate the effect of distance dependence or atom pair composition in the reference state on the atom pair interaction energies, the energies of (A) the C_β atoms of leucines (HH pair) and (B) the C_β atoms of an aspartic acid and a serine (PP pair) were plotted in Figure 3, as a representative for each type of pair. Statistical analyses of these effects were performed in Table II.

First, the effect of the distance-dependent terms in the reference states on the energies was examined. In Figure 3, two black lines (RAPDF in solid and BALL in dashed lines) and two red lines (SHELL in solid and DFIRE in dashed lines) shared the same $p^{\text{expected}}(ab)$, but the two dashed lines (BALL and DFIRE) underestimated $p^{\text{expected}}(r)$ as compared with $p^{\text{observed}}(r)$. These dashed lines were below the solid lines of the same colors for distance $>$7 Å, that is, BALL and DFIRE evaluate the atom pairs at middle range distance more favorably than RAPDF and SHELL, respectively. These propensities were observed both in the C_β atoms of leucines [HH pairs, Fig. 3(A)] and in the C_β atoms of an aspartic acid and a serine [PP pairs, Fig. 3(B)], indicating that the attractive characteristics of BALL and DFIRE, as compared with RAPDF and SHELL, introduced similar changes regardless of the atom pair types.

From Figure 3, we could also observe the effect of the atom pair composition, $p^{\text{expected}}(ab)$, in the reference states on the energies by comparing four (black, red, blue, and green) solid lines. As the fraction of HH pairs in the reference state decreased as RAPDF (black) $>$ SHELL (red) $>$ AKBP (blue) $>$ FAR (green), the energies of the C_β atoms of leucines (HH pair) were decreased [Fig. 3(A)], whereas those of the C_β atoms in an aspartic acid and a serine (PP pair) were increased [Fig. 3(B)]. Thus, the hydrophobic

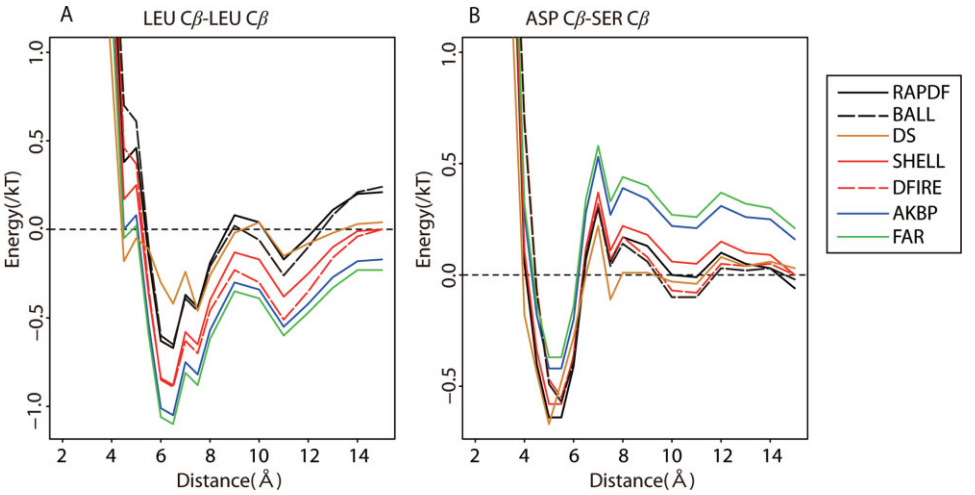

Figure 3. The examples of atom pair interaction energies for (A) an HH pair (C_β atoms of two leucines) and for (B) a PP pair (C_β atoms of an aspartic acid and a serine). The five solid lines share the same distance dependence of reference states, $p^{expected}(r) \equiv p^{observed}(r)$, but have different definitions of $p^{expected}(ab)$. The black, red, blue, green, and orange lines are the RAPDF, SHELL, AKBP, FAR, and DS potentials, respectively. The dashed lines (BALL in black and DFIRE in red) share the same $p^{expected}(ab)$ as the solid lines of the same color (RAPDF in black and SHELL in red, respectively), but they underestimate $p^{expected}(r)$ compared with $p^{observed}(r)$. Statistical summary of these changes in atom pair interaction energies are shown in Table II.

characteristic of the potential exerted opposite effects on HH pairs and PP pairs.

It may be noteworthy that the orange solid line of DS showed peculiar distance dependence as compared with the other solid lines although it shared the same distance-dependent term. It might result from the unique formulation of the reference state, in which the expected atom pair composition depended on the distance (See Table I).

To examine the changes in atom pair interaction energies due to the change in the reference state statistically, the area under the curve within the range 6 Å $< r <$ 15 Å were measured and they were averaged for the HH, HP, and PP pairs (Table II). Underestimating $p^{expected}(r)$ as compared with $p^{observed}(r)$ (RAPDF-BALL and SHELL-DFIRE) decreased the energies equally for all types of atom pairs to exactly the same extent, as indicated by the standard deviation o for these potential pairs. On the other hand, decreasing the fraction of hydrophobic pairs in the reference state (RAPDF-SHELL, RAPDF-AKBP, and RAPDF-FAR) decreased the energies of the HH pairs, but increased those of the PP pairs. These results support the characteristics of the potentials summarized in Table I.

Evaluation of the performances of the potentials

As observed in Figure 3 and Table II, the atom pair interaction energies of the statistical potentials change systematically due to the change in the reference state, rather than fluctuating around some "true" values. Because of these differences, the seven potentials would have different selectivity for the native structures from decoys. The potentials with attractive characteristics would favor compact structures, in which the interatomic distances tend to be short. The potentials with hydrophobic characteristics would favor the clusters of HH pairs, but they might not be able to evaluate favorable hydrophilic interactions, such as salt bridges.

Table II. *Statistical Analyses on the Differences in the Atom Pair Interaction Energies*

	HH	HP	PP
RAPDF−BALL	−0.35 ± 0.00	−0.35 ± 0.00	−0.35 ± 0.00
SHELL−DFIRE	−0.66 ± 0.00	−0.66 ± 0.00	−0.66 ± 0.00
RAPDF−SHELL	−0.62 ± 0.53	0.25 ± 0.56	0.24 ± 0.40
RAPDF−AKBP	−2.33 ± 0.99	0.04 ± 1.21	1.95 ± 1.31
RAPDF−FAR	−2.74 ± 1.19	0.05 ± 1.41	2.22 ± 1.35

The difference in the area under the curve (AUC) of the atom pair interaction energies within the range 6 Å $<$ r $<$ 15 Å between two potentials summarized for HH, HP, and PP pairs.
The first column shows the two compared potentials. The values in the second through fourth columns are the mean and standard deviation of the difference in AUC averaged over HH, HP, and PP pairs.

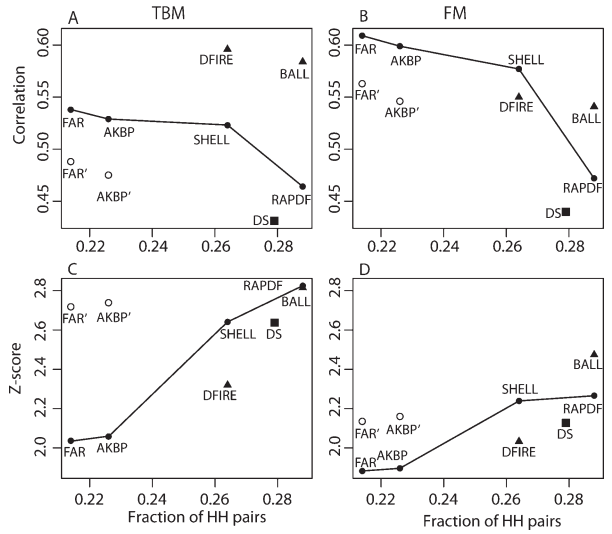

Figure 4. Fraction of hydrophobic atom pairs in the reference state and the performance of potentials measured by Pearson's correlation coefficients (PCC) between the energy and the GDT-TS scores in (A) and (B) and the *Z*-scores of the native structure energies in (C) and (D). PCCs and *Z*-scores were averaged for template-based modeling (TBM) in (A) and (C), and for free modeling (FM) in (B) and (D). Filled circles and filled triangles are the potentials with and without attractive characteristics, respectively. The filled squares represent the DS potential at the average fraction of HH pairs over the distance range of 0–15 Å. The open circles (AKBP′ and FAR′) show the performances of the AKBP and FAR potentials using 3.5–6.5 Å for evaluation.

The effects of the attractiveness and hydrophobicity of the potentials on their performances were assessed for the seven potentials, using the native and model structures in CASP7. The performance was evaluated by correlation coefficients and *Z*-scores for template-based modeling (TBM) and free-modeling (FM) targets (Fig. 4).

We first determined the effect of the hydrophobicity of the four potentials without the attractive characteristic on their performances; that is, RAPDF, SHELL, AKBP, and FAR. Their performance change is depicted by a solid line with filled circles in Figure 4, where the four potentials have lower fractions of hydrophobic pairs from the right to the left, and thus have higher hydrophobic characteristics, as discussed earlier. As seen in the figure, a higher fraction of HH pairs (or lower hydrophobic characteristics) resulted in better performance in the *Z*-scores of native structures, but the performance became worse in the correlation between potential and GDT-TS for both the TBM and FM targets. These results implied that there is a tradeoff between ranking models by structural similarity to the native structure (PCC) and discriminating the native structure from models (*Z*-score). Therefore, higher hydrophobic characteristics of potentials could emphasize the difference among model structures, but prevent the discrimination between the model structures and their native ones. This is possibly because overemphasis on hydrophobic interactions would result in similar potential energies in all compact models. In addition, overemphasis on hydrophobic interactions can result in positive potential energies (or unstable interactions) if the sequence contains many hydrophilic residues. Actually, among the 1788 SCOP native domains, the energies of 98 and 70 domains were evaluated as positive by AKBP and FAR, respectively, where the fraction of charged residues (D, E, R, and K) in the 98 domains with positive potential determined by AKBP was much higher (31%) than that of all the domains (23%).

To lessen the degree of penalties for the hydrophilic residues, the distance range evaluated by these potentials was restricted to 3.5–6.5 Å, according to the proposal by Lu and Skolnick.[15] The open circles in Figure 4 represent the correlations and *Z*-scores of the FAR and AKBP potentials with this distance restriction. We refer to these potentials as AKBP′ and FAR′. They showed a great improvement in the *Z*-score, but their performance decreased in PCC (Fig. 4). These changes in performance were consistent with the idea that the restriction on the interatomic distances for evaluation reduced the hydrophobic characteristics of these potentials.

Next, the effect of attractiveness on performance was examined. BALL and DFIRE displayed attractiveness (favoring compact structures), as compared with RAPDF and SHELL, respectively, as discussed earlier (Table I). For TBM targets, the attractiveness of RAPDF and SHELL greatly increased the correlations, while it decreased the *Z*-score slightly in BALL, as compared with RAPDF, and greatly in DFIRE, as compared with SHELL. On the other hand, for FM targets, the benefit of attractiveness was smaller than that in TBM. The attractive characteristics of potentials may play an important role when model structures have similar folds to the native ones, but they can cause errors for model structures with quite different folds from the native ones.

As depicted in Figure 3, the six potentials tested in this study, except DS, were strongly correlated with each other. Actually, setting RAPDF as the basis potential, the other five potentials can be considered to be the combination of RAPDF with the hydrophobic term (SHELL, AKBP, and FAR), with the attractive term (BALL) or with both the hydrophobic and attractive terms (DFIRE) as characterized in Table I. This interpretation helps us to understand the trade off between *Z*-score and correlation in Figure 4. The energy landscapes of all-atom potentials have been pointed out to be golf-course-like, in which the energy of the native structure is deep below the energies of the model structures, but have only small correlation between energy and the quality of model structures.[19] RAPDF would be one of such golf-course-like potentials. Adding terms that reflect coarse-grained structural features, such as hydrophobicity or attractiveness, to RAPDF resulted in

Table III. *Assessment of the Potentials*

	RAPDF	DFIRE	SHELL	BALL	AKBP	FAR	DS
(A) Template-based modeling targets (104)							
PCC	0.46	0.60	0.52	0.58	0.53	0.54	0.44
Z-score	2.83	2.32	2.64	2.81	2.06	2.03	2.64
ΔGDT	0.075	0.074	0.070	0.053	0.11	0.12	0.087
Enr10	3.03	3.70	3.14	3.81	2.92	2.99	2.94
Enr20	2.212	2.516	2.308	2.540	2.24	2.26	2.16
(B) Free modeling targets (18)							
PCC	0.47	0.55	0.58	0.54	0.60	0.61	0.43
Z-score	2.27	2.03	2.24	2.47	2.16	1.88	2.12
ΔGDT	0.10	0.092	0.10	0.10	0.12	0.11	0.11
Enr10	3.22	2.99	3.78	2.83	3.51	3.49	3.11
Enr20	2.21	2.16	2.48	2.15	2.38	2.47	2.02

Each column shows the average performance of each potential with various evaluation measures for the targets of (A) template-based modeling and (B) free modeling. Enr10 and Enr20 are the 10% and 20% enrichments, respectively. The potential which performed best in each criterion was underlined.

the more funnel-like energy landscape, which would be beneficial in protein structure prediction, but our result suggested that this change was at the cost of decreasing the discrimination of native structure.

The reference state of DS was unique from other reference states, in the definition of the expected probability of atom pair *ab* at a distance *r* (Table I). This reference state considered the numbers of atom pairs at a distance *r* in which atom *a* or *b* was involved, $N_{ax}^{\text{observed}}(r)$ or $N_{bx}^{\text{observed}}(r)$. In this reference state, $p^{\text{expected}}(ab)$ depended on the distance *r*; that is, $p^{\text{expected}}(ab|r)$. The expectation of an atom pair *ab* in the reference state of DS became large if atom types *a* and *b* belonged to hydrophobic atom pairs, which was similar to the reference state of RAPDF, in terms of hydrophobic characteristics. The correlation and the *Z*-score of DS are a little worse than those of RAPDF. This weakness of DS might result from the fact that its reference state was more complicated, and it may suffer from the noise in the statistics.

The performance of the seven potentials measured in the five assessment criteria (see Materials and Methods) were summarized in Table III (a) for TBM targets and (b) for FM targets. As demonstrated by the ΔGDT values, among the seven potentials, BALL and DFIRE selected the models with the least error for TBM targets and for FM targets, respectively. This may imply the importance of selecting compact models by attractive potential.

However, the enrichment criteria showed the opposing trend for TBM and FM targets. For TBM targets, the enrichment scores were better in DFIRE and BALL than those in other five potentials, but for FM targets, these two potentials performed poorly. The enrichment results support the larger benefit of attractive potentials where the differences between the model structures and the native structure are small.

The seven potentials tested here had their advantages and disadvantages, and that it was impossible to define the best reference state in a practical sense as well as in a theoretical sense.

Materials and Methods

Protein domain sets to calculate the statistical potentials

From all domains (9536) in the SCOP40 databases,[20,21] we eliminated the following structures: not determined by X-ray crystallography (1440), with >3.5 Å resolution (35), classified as membrane proteins (441), with missing residues other than N or C termini (2671), with too few (<50 residues) or too many (>800 residues) residues (245), with other domains in the same chain (2694) and with deformed spherical forms (222). The degree of deformation from a sphere of a domain was defined by the ratio of the smallest radius of inertia calculated for all its Cα atoms to the largest one. When the degree is smaller than 0.4, the domain was judged as a deformed domain. Finally, we obtained 1788 representative domains, and they were used to construct all of the statistical potentials in this study.

Calculation details of $N_{ab}^{\text{observed}}(r)$

We considered 167 atom types, by treating all of the nonhydrogen atoms as having different atom types when they are in different amino acid residues.[16] All of the atom pairs in a structure, except those that belong to the same residue or to the neighboring residues on sequence, were considered. We used the distance intervals described by Zhou and Zhou[14] to count the frequency at each distance. The width of the bin was 2 Å for $r < 2$ Å, 0.5 Å for $2 < r < 8$ Å, and 1 Å for $8 < r < 15$ Å, and the total number of bins was 20. The atom pairs with more than 15 Å separation were not considered in the calculation of potentials.

Performance evaluation of each potential

The performance of the potentials was evaluated using the decoy sets of CASP (critical assessment on the techniques for protein structure prediction) experiments. A decoy set is composed of the native structure and hundreds of model structures for the same amino

acid sequence. These decoy sets included 122 targets in CASP7.[1] All of the native and model structures were obtained via the CASP web site. The quality of a model structure was defined by its structural similarity to the native structure, the GDT-TS score.[22] GDT-TS is the average of the fraction of matched residues within various distance thresholds after global superposition. This score was scaled within [0,1], and thus it approached 1 as the model structure became close to the native structure.

Five assessment criteria were used: (1) PCC, the Pearson's correlation coefficient between the energy and the GDT-TS score of the structure; (2) *Z*-score, the difference between the native energy and the average of the decoy energies in standard deviation units; (3) ΔGDT, the difference in the GDT-TS scores between the best-scored model and the best model; and (4, 5), 10% and 20% enrichments; *x*% enrichment is the relative frequency of the most accurate *x*% models in the GDT-TS score among the *x*% best scored models and *x*%. Therefore, the maximum value of the *x*% enrichment is 1.0/(*x*/100). These five assessment measures were calculated for each target domain of CASP7, using each of the potentials or scores, and they were averaged for 104 TBM (template-based modeling) targets and 18 FM (free modeling) targets.

Since Pearson's correlation coefficients and *Z*-scores assume the normal distributions of the variables, only the model structures with GDT-TS scores better than a threshold were used. This threshold was determined for each target in the following manner. First, the occurrence probabilities of the GDT-TS scores were counted for each 0.05 of GDT-TS. The probabilities were accumulated from the highest GDT-TS bin in descending order, until the accumulative probability exceeded 0.65 (i.e., enough structures had been accumulated) and the last bin was searched until its probability became less than 0.05, to ensure that the last bin was not in another peak. The floor of the last bin was set to be the threshold of the GDT-TS score for the target. This treatment enabled us to focus on the evaluation of relatively good models. On the other hand, because ΔGDT or enrichments do not assume the normal distribution, all of the model structures were used for the assessment.

Conclusions

We separated the reference states of all-atom distance-dependent potentials into the distance dependence term and the atom pair composition term, and formulated the existing four potentials and developed three new potentials. The difference in the distance ranges for monitoring the atom pair composition in the reference state introduced the changes in hydrophobic characteristic of the potentials. The hydrophobic characteristic improved the correlation between the energy and the GDT-TS score, but decreased the *Z*-score of the native structure. By underestimating the probability of atom pairs in the reference state in the middle range (7 Å $< r <$ 14 Å), the potential with attractive characteristic could be obtained. This attractive characteristic was beneficial in structure evaluation because it reflects the globularity of the proteins, but it sometimes caused errors in FM targets.

Our research focused on the factors in all-atom statistical potentials that have not been explored well, that is, attractiveness and hydrophobicity. After examining several potentials with different attractiveness and hydrophobicity, we conclude that there would be no best reference state on these characteristics that can improve both in ranking of model structures and in discriminating the native structure. This result would be the realization of the trade off between the golf-course-like and funnel-like energy landscapes, the problem to which many of the potentials are confronting.

Acknowledgments

Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

1. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A (2007) Critical assessment of methods of protein structure prediction-Round VII. Proteins 69 (Suppl 8):3–9.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242.
3. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170.
4. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362.
5. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256:623–644.
6. Simons KT, Kooperberg C, Huang E, 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.
7. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15: 2507–2524.
8. Xia Y, Huang ES, Levitt M, Samudrala R (2000) Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 300:171–185.
9. Zhou H, Zhou Y (2004) Quantifying the effect of burial of amino acid residues on protein stability. Proteins 54: 315–322.
10. Yang Y, Zhou Y (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.
11. Lu H, Lu L, Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. Biophys J 84:1895–1901.
12. Sippl MJ (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.

13. Rykunov D, Fiser A (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins 67: 559–568.
14. Zhou H, 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.
15. Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 44:223–232.
16. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275:895–916.
17. Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they?J Mol Biol 257:457–469.
18. DeBolt SE, Skolnick J (1996) Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng 9: 637–655.
19. Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol 16: 166–171.
20. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
21. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192.
22. Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31:3370–3374.