ACCOUNTS
of chemical research

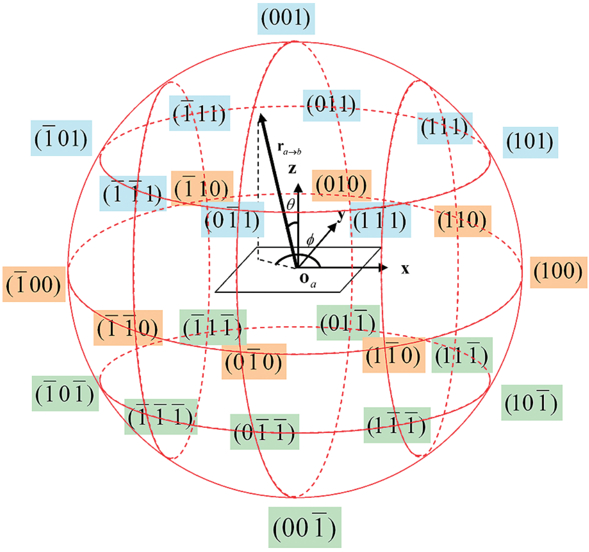
Explicit Orientation Dependence in Empirical Potentials and Its Significance to Side-Chain Modeling

JIANPENG MA*

Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, and Department of Bioengineering, Rice University, Houston, Texas 77005

RECEIVED ON JANUARY 7, 2009

CONSPECTUS



Protein structure modeling and prediction have important applications throughout the biological sciences, from the design of pharmaceuticals to the elucidation of enzyme mechanisms. At the core of most protein modeling is an energy function, the minimum of which represents the free energy “cost” for forming a correct protein structure. The most commonly used energy functions are knowledge-based statistical potential functions; that is, they are empirically derived from statistical analysis of a set of high-resolution protein structures. When that kind of potential function is constructed, the anisotropic orientation dependence between the interacting groups is a critical component for accurately representing key molecular interactions, such as those involved in protein side-chain packing. In the literature, however, many potential functions are limited in their ability to describe orientation dependence. In all-atom potentials, they typically ignore heterogeneous chemical-bond connectivity. In coarse-grained potentials, such as (semi)-residue-based potentials, the simplified representation of residues often reduces the sensitivity of the potential to side-chain orientation.

Recently, in an effort to maximally capture the orientation dependence in side-chain interactions, a new type of all-atom statistical potential was developed: OPUS-PSP (potential derived from side-chain packing). The key feature of this potential is its explicit description of orientation dependence in molecular interactions, which is achieved with a basis set of 19 rigid-body blocks extracted from the chemical structures of 20 amino acid residues. This basis set is specifically designed to maximally capture the essential elements of orientation dependence in molecular packing interactions. The potential is constructed from the orientation-specific packing statistics of pairs of those blocks in a nonredundant structural database. On decoy set tests, OPUS-PSP significantly outperforms most of the existing knowledge-based potentials in terms of both its ability to recognize native structures and its consistency in achieving high *Z* scores across decoy sets. The application of OPUS-PSP to conformational modeling of side chains has led to another method, called OPUS-Rota. In terms of combined speed and accuracy, OPUS-Rota outperforms all of the other methods in modeling side-chain conformation.

In this Account, we briefly outline the basic scheme of the OPUS-PSP potential and its application to side-chain modeling via OPUS-Rota. Future perspectives on the modeling of orientation dependence are also discussed. The computer programs for OPUS-PSP and OPUS-Rota can be downloaded at http://sigler.bioch.bcm.tmc.edu/MaLab. They are free for academic users.

Published on the Web 05/15/2009 www.pubs.acs.org/acr
10.1021/ar900009e CCC: $71.50

Introduction

Knowledge-based statistical energy functions are widely used in protein structure modeling and prediction.[1] They are usually constructed on the basis of statistical analysis of predefined interacting units from a set of selected high-resolution structures. The interacting units can be either coarse-grained structural components, such as Cα atoms for representing a whole residue, or atomistic structural components, as in all-atom representation. The energy function is the potential of the mean force or free-energy cost, required for generating the observed distribution of the interacting units in the real structures from a zero-interaction reference state. Thus, the choices of interacting units are crucial for the effectiveness of the energy functions. One of the key issues is the orientation dependence in the interaction between the units. This is because the chemical bond connectivity is often ignored in constructing statistical energy functions, leading to mis- or under-representation of anisotropic orientation preference in molecular interactions.

In the literature, substantial efforts have been made to model anisotropic orientation preference.[2−9] An early attempt employed a side-chain-specific local reference frame to construct distance- and orientation-dependent residue-based statistical potentials for proteins.[10] In a subsequent work,[4] it was shown that contacts between side chains and main chains are important and a Cα-SC-Pep model was introduced to represent orientation dependence. In a more recent highly coarse-grained potential, called OPUS-Ca,[8] the orientation preference was introduced into a distance-dependent pairwise potential. In that case, the orientation dependence between two side chains was described by the relative orientation between two Cα−Cβ vectors. It was found that inclusion of this effect improved the ability of the potential to recognize the native state and to improve *Z* scores in decoy set tests. Orientation dependence for homodimeric[11] and heterodimeric[12] interactions among seven hydrophobic residues in water has also been included in an analytical modeling of potentials of mean force.

Although a certain degree of success in describing orientation dependence was achieved in the aforementioned work, there is still much room for improvement. Recently, a new type of potential, called OPUS-PSP, was developed to maximally capture the orientation dependence in side-chain interactions.[13] OPUS-PSP is an orientation-dependent statistical all-atom potential derived from side-chain packing.

Here, we first briefly outline the general framework of OPUS-PSP, followed by the results of its performance on decoy set tests. Then, we will discuss a major application of OPUS-PSP on side-chain conformation modeling via a method called OPUS-Rota.[14] Most importantly, on the basis of the lessons learned from our own work and others, we will discuss issues and insights in the modeling of orientation dependence in molecular interactions.

Theoretical Framework of OPUS-PSP

OPUS-PSP is constructed from two major components: (a) a novel set of 19 rigid-body blocks that define the geometry of the interaction units and (b) a knowledge-based energy function based on packing statistics of these blocks. In addition, a repulsive Lennard−Jones (LJ) term is used to deter steric clashes. Coarse-graining and symmetry are also employed to improve the statistics.

Definitions of Rigid-Body Blocks and Relative Orientation. First, to form the basis set of interaction units, the chemical structures of 20 residues are decomposed into a set of 19 rigid-body blocks (shown in Figure 1a). Those blocks share three important characteristics: (a) all atoms in a block are chemically bonded and belong to the same residue; (b) each block is treated as a rigid body; (c) all non-hydrogen heavy atoms are assumed to be in the same plane. For the proline ring of block type 19, assumptions b and c are approximate and we found that they are reasonable in constructing OPUS-PSP. Furthermore, the α carbon atoms of all residues, except Pro and Gly, are not included in the basis set. We do so by assuming that the heavily shielded α carbons have minimal influence on side-chain packing, and our results support this assumption. In this representation, each residue contains more than one block but each block appears only once in a single residue. Figure 1b shows the block compositions of the 20 residue types. For notational consistency, we shall denote residue types (20 total) with *m* and *n*, block types (19 total) with *a* and *b*, block indices with α and β, and atomic indices with *i* and *j*.

A special coordinate system is designed to define the relative orientation of a pair of blocks. As illustrated in Figure 2, the relative orientation of block types *a* and *b* is defined using three variables: two relative direction vectors $\mathbf{r}_{a\to b}$ and $\mathbf{r}_{b\to a}$ and an inter-rotation angle ψ_{ab} along the axis connecting the origins of the two blocks in their respective molecular reference frames. These coordinates describe the axial rotation around the line linking the origins of the two blocks and the pivot motion around the origin of each block, respectively. The relative orientation of a pair of blocks is completely defined by

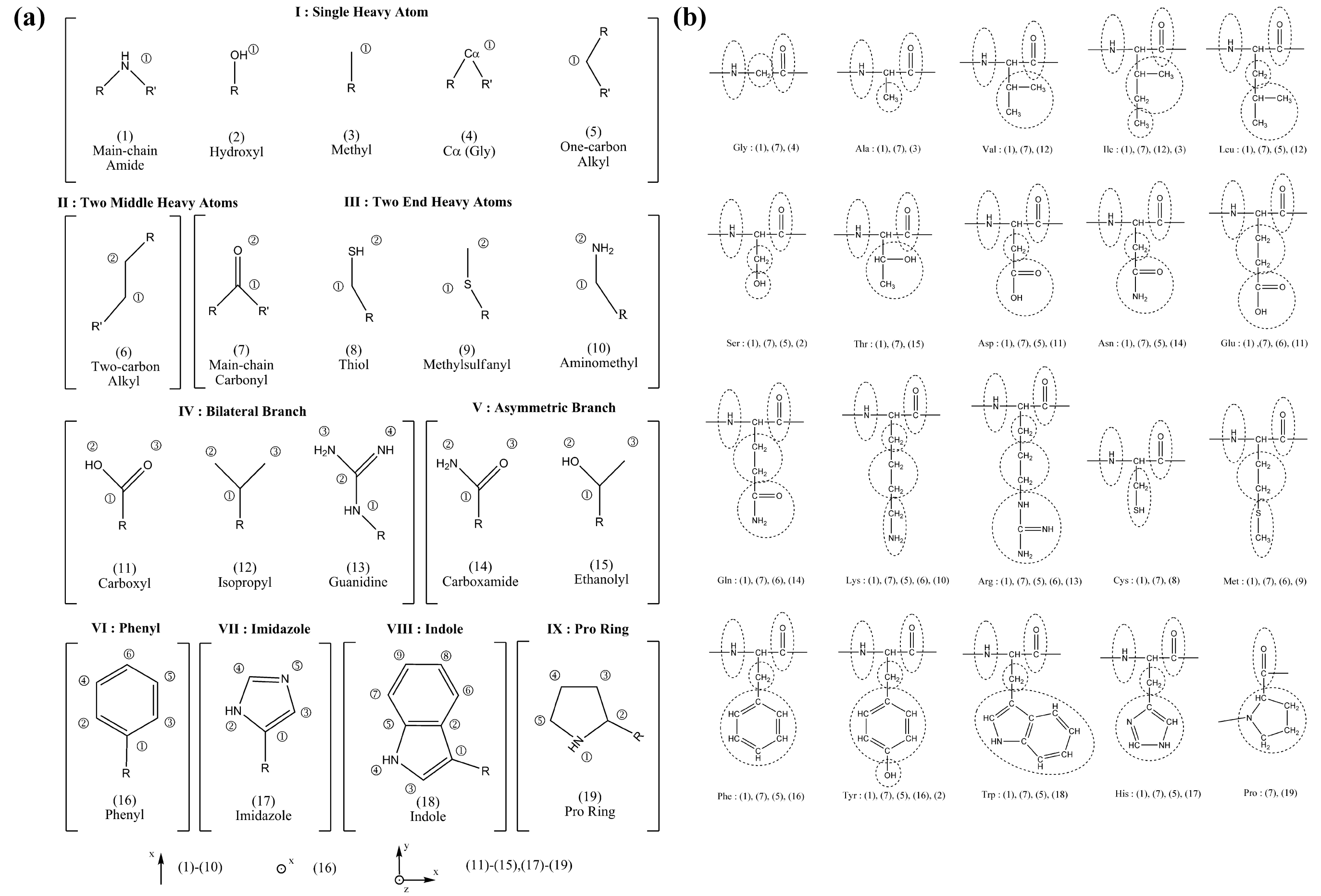


FIGURE 1. Rigid-body blocks in OPUS-PSP. (a) Definition of 19 block types. Blocks are categorized into nine symmetry classes denoted by Roman numerals. Block classes I, II, III, and VI are line shapes, and the others are plane shapes. R and R′ are not considered parts of the blocks but are shown to indicate connectivity only. The reference frames for line and plane shapes are schematically shown alongside their corresponding block types at the bottom of the figure. (b) Block composition of residues. All blocks (block types denoted by numbers in parentheses and defined in Figure 1a) are circled for all amino acids. This figure is adopted from Figure 1 in ref 13.

these three variables (computed in the laboratory reference frame), coupled with the molecular reference frame for each block.

Energy Function. OPUS-PSP contains an orientation-dependent packing energy term E_{orient} and a repulsive energy term E_{repul}

$$E_{\text{PSP}} = E_{\text{orient}} + w_{\text{repul}}E_{\text{repul}} \tag{1}$$

where w_{repul} is a weight parameter optimized against a small subset of decoy sets.[13]

To calculate the first term, the total orientation-dependent packing energy, E_{orient}, we first define the packing energy for a pair of blocks by

$$E(\Omega_{ab}, a,b) = -k_{\text{B}}T \log \frac{p^{\text{obs}}(\Omega_{ab}, a,b)}{p^{\text{ref}}(\Omega_{ab}, a,b)} \tag{2}$$

Here, p^{obs} is the probability of a particular orientation state for block types a and b in contact with respect to all observed contact states for any block pair extracted from the nonredundant structure database, and p^{ref} is the contact probability of all possible occurrences of that state without packing interactions (the reference state). The quantity $\Omega_{ab} = (\mathbf{r}_{a\to b}, \mathbf{r}_{b\to a}, \psi_{ab})$ designates the relative orientation of a and b, and $k_{\text{B}}T$ is the Boltzmann constant (set to unity). The value of E_{orient} is obtained by summing the packing energies of all pairs of blocks in contact ("block contact pairs") between all pairs of nonconsecutive residues

$$E_{\text{orient}} = \sum_{\alpha,\beta} \delta(\alpha,\beta)\hat{E}(B(\alpha),B(\beta)) \tag{3}$$

Here, $\delta(\alpha,\beta)$ is a delta function, whose value is 1 when blocks α and β are in contact and 0 otherwise, and $B(\alpha) = a$ maps block α to its block type a. The second term in eq 3 is $\hat{E}(a,b) = n(a,b)E(\Omega_{ab}, a,b)$, where $n(a,b)$ is a weighting term for block size defined as the average number of pairs of heavy atoms in contact between block types a and b (we define an "atom

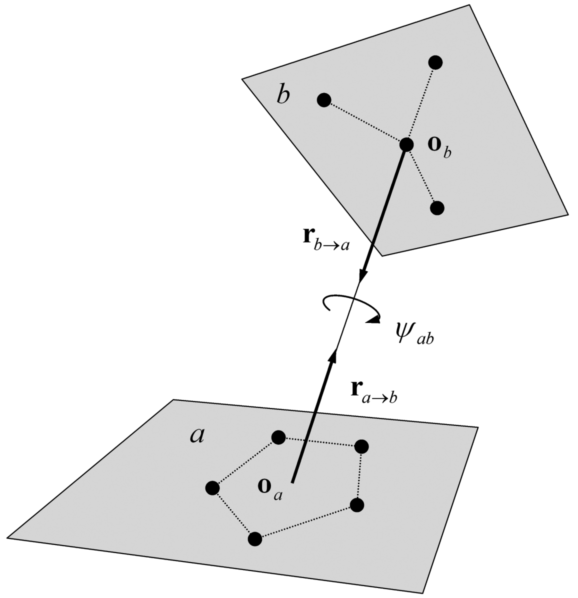


FIGURE 2. Definition of the relative orientation of blocks in OPUS-PSP. If block types *a* and *b* are in contact, then $\mathbf{r}_{a\to b}$ and $\mathbf{r}_{b\to a}$ are the relative direction vectors and ψ_{ab} is the inter-rotation angle along the axis connecting the origins $\mathbf{o}_a$ and $\mathbf{o}_b$ of the two blocks. This figure is adopted from Figure 2 in ref 13.

contact pair" as two atoms whose pairwise distance is less than 5 Å). The weighting term is evaluated by random sampling in the manner of the reference state probability calculation. This is necessary because larger blocks contribute more atom contact pairs and therefore more energy. In calculating E_{orient}, the contribution is restricted to side-chain–side-chain and main-chain–side-chain interactions only. The main-chain–main-chain hydrogen bonding and other short-range interactions are not included.

The repulsive term E_{repul} is defined as

$$E_{\text{repul}} = \sum_{i,j} E_{\text{LJ}}(i,j) \tag{4}$$

where $E_{\text{LJ}}(i,j)$ is a repulsive (no attractive term) LJ potential for two atoms *i* and *j*. Similar to E_{orient}, the summation in the LJ term ignores interactions between pairs of main-chain atoms and between two atoms in the same residue. Note that E_{orient} and E_{repul} are typically orthogonal; therefore, overcounting is not an issue.

Coarse Graining of Orientation Bins and Symmetry. It is necessary to coarse grain the orientation space and exploit the symmetry of the 19 blocks given the limited amount of nonhomologous protein data available. As shown in Figure 1a, these blocks are classified into nine symmetry classes that belong to two basic groups: plane shapes (IV, V, and VII–IX) and line shapes (I–III and VI). Note that VI is regarded as a line shape because of the 6-fold axial symmetry of the phenyl ring.

For each plane-shaped block, the relative direction with respect to the molecular reference frame of the block is coarse-grained into 26 bins (illustrated in Figure 3a). For each line-shaped block, the cylindrical symmetry allows usage of five latitudinal bins (shown in Figure 3b). Figure 3c describes the θ and ϕ ranges of each relative direction bin. The inter-rotation angle is coarse-grained into four bins spanning $\pi/2$ radians each. In our study, we found that a choice of 26 directional bins is appropriate for plane-shaped blocks to balance the trade-off between the number of bins and the available structure data for statistical analysis.

For two blocks in contact, the maximal number of bins is $26 \times 4 \times 26 = 2704$. However, in practice, certain redundant bins are consolidated on the basis of the intrinsic molecular symmetry of the blocks. This leads to a much smaller number of bins.

Performance of OPUS-PSP on Decoy Set Recognition

The performance of OPUS-PSP was examined in benchmark studies using the popular decoy set collections: Decoys 'R' Us,[15] HR,[16] Rosetta (and Rosetta2),[17,18] MOULDER,[19] structal (http://dd.compbio.washington.edu/), and the decoy sets collected by Gilis,[20] which we call the Gilis collection. The results are presented in Table 1. Of all of the benchmarks, only the MM-PBSA[21] and MJ_2005 potentials[7] outperformed OPUS-PSP on the structal decoy sets. These decoy sets contain decoys generated by comparative modeling of globins and immunoglobulins [60% of them have a Cα root-mean-square deviation (rmsd) less than 2.5 Å from the native conformation]. For the ig_structal and ig_structal_hires sets, OPUS-PSP can do better if main-chain interactions between pairs of block types {1,5,6,7} are also included in the total energy calculation.

OPUS-Rota: A Fast and Accurate Method for Side-Chain Modeling

Side-chain conformation modeling is one of the most severe bottlenecks in the high-accuracy refinement of computationally predicted structures. Aided by OPUS-PSP, OPUS-Rota[14] is a new method developed for such a purpose.

Rotamer libraries are most commonly and successfully used by side-chain modeling methods to reduce the space of conformations that must be sampled, and there are many rotamer-based side-chain modeling method, as summarized in the OPUS-Rota paper.[14] In the rotamer approach, side-chain conformations are limited to a small set of most likely posi-

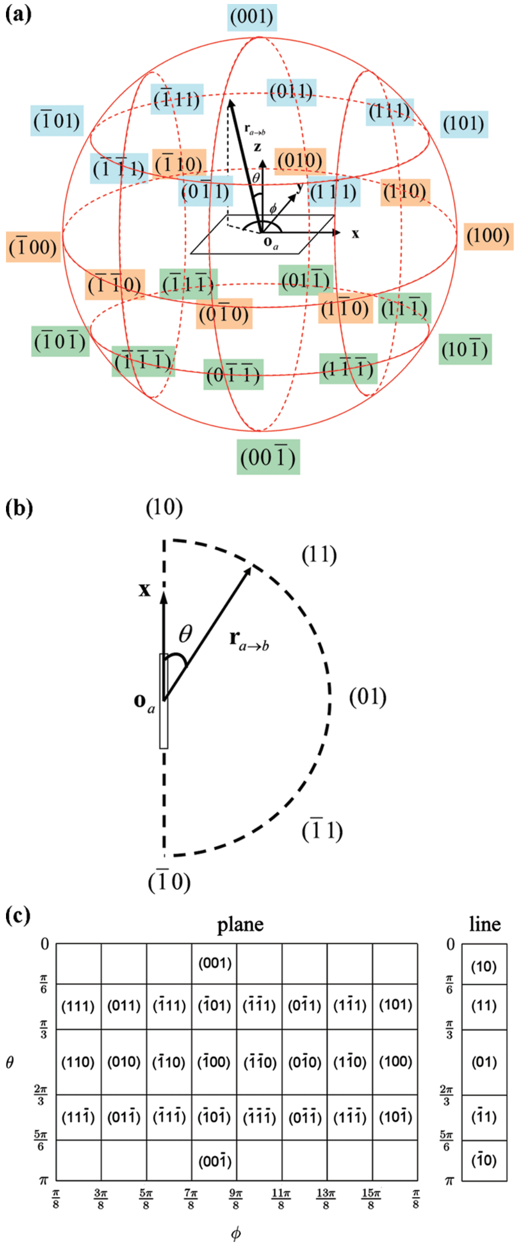


FIGURE 3. Definition of the relative direction bins for line- and plane-shaped blocks in OPUS-PSP. (a) A total of 26 relative direction bins for plane-shaped blocks (classes IV, V, and VII−IX). Each bin is denoted by the index ($n_xn_yn_z$) and is derived from the spherical angles θ and ϕ of vector $\mathbf{r}_{a\to b}$ in the reference frame of block *a*. (b) A total of 5 relative direction bins for line-shaped blocks (classes I−III and VI). Each bin is denoted by the index (n_xn_y) and is derived from the angle θ between the primary axis (*x* axis) and vector $\mathbf{r}_{a\to b}$ formed from the origin $\mathbf{o}_a$ of block *a* to the origin $\mathbf{o}_b$ of block *b*. (c) Direction bin indices plotted on a Mercator projection, for illustration only (a Mercator projection is a cylindrical map projection and the most common geographic map projection). The ranges for spherical angles θ and ϕ are indicated on the axes of the map. For plane shapes, the first or last row of the map represents a single bin at each of the poles rather than eight individual cells. The 5 bins for line shapes (on the right) are consolidated from the 26 latitudinal bins of the plane shapes. This figure is adopted from Figure 3 in ref 13.

TABLE 1. OPUS-PSP Performance on Various Decoy Sets[a]

(a) OPUS-PSP Performance Compared to Other Potentials

	top 1/total number[b]	mean *Z*
	Decoys ‘R’ Us[18,45−48]	
OPUS-PSP	31/34	−5.37
HPMF[49]	29/32[c]	−4.18
DOPE[39]	28/32	
MSE[50]	21/23	−5.78
DFIRE[38]	27/32	−4.52
MJ_2005[7]	27/34	−5.93
DFIRE-SCM[51]	23/32	−4.36
MM-PBSA[21]	23/34	−1.95
DGR[52]	21/25	−5.25
DWL[53]	21/32	−3.66
TE13[54]	14/25	−3.53
CALSP[55]	15/25	
Rosetta[6,18,56]	14/32[d]	
	MOULDER[19]	
OPUS-PSP	19/20	−4.60
DOPE	19/20[d]	
Rosetta	19/20[d]	
DFIRE	19/20[d]	
DFIRE-SCM	19/20[d]	
	HR[16]	
OPUS-PSP	135/148	−7.50
HR[16]	113/150	
TE13	92/148[e]	
	Rosetta (X-ray)[18]	
OPUS-PSP	37/41	−6.56
DFIRE	31/41	−3.91
DFIRE-SCM	33/41	−4.90
CALSP	28/41	−4.16
	Rosetta2[17,18]	
OPUS-PSP	23/41	−2.71
OPUS-PSP (X-ray)	22/25	−4.49
DOPE	11/41[f]	−1.50
	Rosetta 1 + 2[g] (X-ray)[17,18]	
OPUS-PSP	34/35	−6.76
HPMF	30/35	−4.42
	hg_structal[h]	
OPUS-PSP	18/29	−1.76
MM-PBSA	20/29	−1.60
MJ_2005	22/29	−2.76
	ig_structal[h]	
OPUS-PSP	46/61[i]	−2.79
MJ_2005	49/61	−3.55
	ig_structal_hires[h]	
OPUS-PSP	19/20[i]	−3.03
MJ_2005	19/20	−4.31
	Gilis[20]	
OPUS-PSP	43/45	−5.58

(b) OPUS-PSP Performance on Decoys ‘R’ Us

	PDB code	decoy set size	rank	*Z* score
		4state_reduced		
1	1ctf	631	1	−4.23
2	1r69	676	1	−4.52
3	1sn3	661	1	−5.35
4	2cro	675	1	−3.77
5	3icb	654	1	−2.72
6	4pti	688	1	−5.97
7	4rxn	678	1	−4.32
		fisa		
8	1fc2	501	312	0.25
9	1hdd-C	501	1	−4.10
10	2cro	501	1	−5.05
11	4icb	501	1	−7.40

TABLE 1. Continued

(b) OPUS-PSP Performance on Decoys 'R' Us

	PDB code	decoy set size	rank	*Z* score
		fisa_casp3		
12	1bg8-A	1201	1	−6.01
13	1bl0	972	1	−6.00
14	1eh2	2414	1	−4.42
15	1jwe	1408	1	−7.95
16	smd3	1201	1	−6.73
		lattice_ssfit		
17	1beo	2001	1	−9.58
18	1ctf	2001	1	−6.78
19	1dkt-A	2001	1	−6.75
20	1fca	2001	1	−6.13
21	1nkl	2001	1	−4.40
22	1pgb	2001	1	−7.79
23	1trl-A	2001	1	−4.81
24	4icb	2001	1	−5.95
		lmds		
25	1b0n-B	498	1	−4.74
26	1bba	501	501	3.66
27	1ctf	498	1	−8.99
28	1dtk	216	1	−6.07
29	1fc2	501	409	0.94
30	1igd	501	1	−7.77
31	1shf-A	438	1	−7.87
32	2cro	501	1	−7.17
33	2ovo	348	1	−5.87
34	4pti	344	1	−8.15

[a] This table is adopted from Table 1 in the original OPUS-PSP paper.[13] [b] "total number" is the total number of decoy sets used for a specific decoy set collection, and this number may vary from study to study in the literature even for the same collection. [c] OPUS-PSP recognizes 30 of the 32 decoy sets used for HPMF. [d] Results taken from ref 39. [e] Results taken from ref 16. [f] Results taken from ref 57. [g] The total number of 35 is a subset of X-ray structures in the combined Rosetta and Rosetta2 collections. [h] From http://dd.compbio.washington.edu/. [i] OPUS-PSP includes main-chain interactions of block types {1,5,6,7}.

tions (rotamers) taken from a rotamer library derived from X-ray structures.

Fast rotamer methods, such as SCWRL,[22] can quickly locate the global minimum by using a simple pairwise energy function and dead-end elimination (DEE).[23,24] The accuracy of such methods is limited because the energy function used is oversimplified.[25,26] Methods that use more accurate energy functions, such as NCN[27] and LGA,[28] are significantly slower because of computationally expensive long-range and multibody terms. High computational cost limits the application of these methods because the speed of execution in side-chain modeling is very important in the iterative process of structure prediction.

Brief Outline of the OPUS-Rota Algorithm. The total energy function used in OPUS-Rota has four terms

$$E_{\text{total}} = w_{\text{orient}}E_{\text{orient}} + w_{\text{vdw}}E_{\text{vdw}} + E_{\text{rot}} + w_{\text{solvation}}E_{\text{solvation}} \tag{5}$$

Here, E_{orient} is the side-chain packing potential OPUS-PSP,[13] which is a short-range, pairwise, and coarse-grained all-atom potential that allows for fast and accurate energy evaluation during intensive sampling. The second term E_{vdw} is a modified 6−12 LJ potential also used in OPUS-PSP; E_{rot} is a term related to rotamer frequency; and $E_{\text{solvation}}$ is a solvation energy term. The three weights, $w_{\text{orient}} = 0.15$, $w_{\text{vdw}} = 1.0$, and $w_{\text{solvation}} = 0.1$, are obtained by optimizing against a small set of high-resolution structures.

The third rotamer frequency term E_{rot} has the same form used in SCWRL.[22] However, the contributions of bulky ring side chains {Phe, Tyr, Trp, or His} are scaled up by a factor of 3. The rotamer frequencies are taken from Dunbrack's rotamer library.[29]

Similar to what was used in the literature,[30] the solvation energy $E_{\text{solvation}}$ takes the form

$$E_{\text{solvation}} = \sum_i \Delta\sigma_i S_i \tag{6}$$

where S_i is the solvent-accessible surface area (SASA) of atom i and $\Delta\sigma_i$ is the atomic solvent parameter from Sharp et al.[31] To rapidly calculate SASA, OPUS-Rota adopts the pairwise approximation method of Zhang et al.[32]

OPUS-Rota uses simulated annealing by heat-bath Monte Carlo as a sampling method,[33] which is able to rapidly identify near-native conformations when combined with neighbor list techniques and efficient energy updates. In OPUS-Rota, the move set for a given main-chain conformation is the collection of rotamer states from Dunbrack's rotamer library,[29] selected in order of highest to lowest probability until the cumulative probability reaches at least 99.5%. In this way, almost all possible rotamers can be sampled.

Performance of OPUS-Rota. The performance of OPUS-Rota was benchmarked with 65 high-resolution X-ray structures used in the literature.[27,34] The analysis was carried out for both overall (all residues) and core residues. Core residues are defined as residues with a solvent-accessible ratio below 17% (53.5% of residues are found to be core residues by this definition). The accuracy of χ_1 is defined as the percentage of residues whose predicted χ_1 dihedral is no more than 40° from the native value. The accuracy of χ_{1+2} is defined as the percentage of residues for which both χ_1 and χ_2 are in the 40° range.

Figure 4 shows the accuracy of OPUS-Rota for each residue type. Serine has the lowest χ_1 accuracy for all residues and core residues. Polar and charged residues have lower χ_{1+2} accuracy, especially flexible surface residues. Hydrophobic and aromatic residues consistently have high accuracy, except for His, which has high χ_1 accuracy (overall, ~93%) but low χ_{1+2}

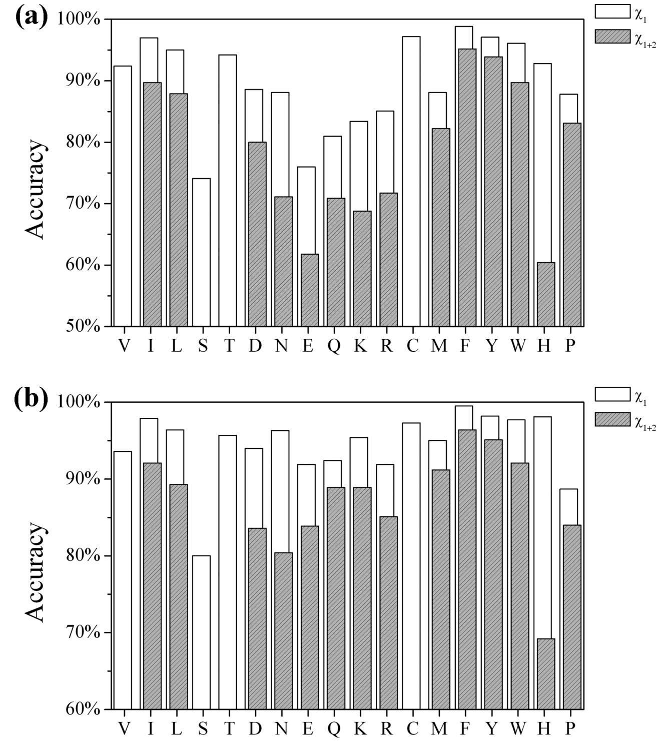


FIGURE 4. Accuracy of OPUS-Rota for each residue type. (a) Overall χ_1 and χ_{1+2} accuracies. (b) Core residue χ_1 and χ_{1+2} accuracies (core residues are defined as the residues whose solvent-accessible ratio is below a cutoff of 17%). This figure is adopted from Figure 2 in ref 14.

TABLE 2. Accuracy and Speed of OPUS-Rota and Several Other Side-Chain Modeling Methods on the 65-Protein Test Set[a]

	all residues		core residues[b]			
	χ_1 (%)	χ_{1+2} (%)	χ_1 (%)	χ_{1+2} (%)	execution time	references
OPUS-Rota	89.0	79.1	94.5	88.7	9.6 min[d]	
SCWRL	83.6	70.3	88.8	79.2	2.2 min + 5 h[c,d]	22
NCN	89.3	77.5	94.1	87.4	24 h[f]	27
LGA	88.5	74.1	93.7	84.6	14 h[f]	28
SPRUCE	86.7	74.0	93.7	86.7	20 h[e]	34
Rosetta	85.1	72.7	91.5	84.5	43.7 h[d]	58
$\text{SCAP}_{\text{orig}}$[g]	84.1	70.7	90.7	82.5	2.1 h[d]	25
$\text{SCAP}_{\text{modi}}$[g]	83.1	70.1	91.4	84.0	24 h[f]	27

[a] This table is adopted from Table 2 in the original OPUS-Rota paper.[14] [b] Tests on OPUS-Rota, SCWRL, SPRUCE, Rosetta, and $\text{SCAP}_{\text{orig}}$ use the same definition of core residues (SPRUCE uses different solvent parameters and a different cutoff), while NCN, LGA, and $\text{SCAP}_{\text{modi}}$ define the core as having <20% accessible surface area in the native structure, according to the method by Lee and Richards.[59] All of the definitions result in a similar portion of core residues, ∼53.5%.[34] [c] SCWRL requires >5 h for protein 1qlw but only 2.2 min for the remaining 64 proteins. [d] Times for OPUS-Rota, SCWRL, Rosetta, and $\text{SCAP}_{\text{orig}}$ are for a single run on one Intel Xeon 2.8 GHz processor (by the software provided by the authors). [e] SPRUCE is run on one Intel Xeon 3.2 GHz processor.[34] [f] Data for run times are from ref 27. [g] $\text{SCAP}_{\text{orig}}$ is the original version of SCAP[25] (executable provided by the authors), and $\text{SCAP}_{\text{modi}}$ is the modified version of SCAP from ref 27, in which a larger rotamer library is used.

accuracy (overall, ∼60%; core, ∼70%). This is probably due to the lack of knowledge of protonation states.

OPUS-Rota outperforms other related methods in terms of combined speed and accuracy. As shown in Table 2, on the 65-protein test set mentioned above, OPUS-Rota is much faster than all other methods except SCWRL,[22] which is similar in speed. In addition, OPUS-Rota is much more accurate than SCRWL and comparably accurate with the rest. The computational efficiency of OPUS-Rota scales linearly with protein size.

For real applications in structure prediction, both SCWRL and OPUS-Rota were also tested on the Wallner and Elofsson homology modeling benchmark set.[35] It was found that OPUS-Rota performs consistently better than SCWRL when sequence identity is higher than 40% (see Figure 3 in ref 14). When sequence identity is lower than 40%, both methods have low accuracy, which is an expected result because the template structures are so far away from the target structures. This indicates that the quality of side-chain modeling heavily depends upon the accuracy of the main-chain coordinates.

Discussion and Future Perspective

The most important feature of OPUS-PSP is its unique basis set of 19 rigid-body blocks that captures the essential elements of anisotropic orientation-dependent molecular interactions. OPUS-PSP is designed to maximally sense the change of relative orientation between two packed blocks, even when there is insignificant change in the packing distance. To the best of our knowledge, this is a feature that no other potential possesses.

OPUS-PSP is not a distance-dependent potential. The effect of packing distance between atoms is implicitly contained in its form. For example, if two blocks are in contact with native packing orientation, then the atomic contact criteria used in OPUS-PSP and the orientation parameters will restrict the distances between the atoms because of the fixed sizes of the blocks.

OPUS-PSP does not model solvation effects explicitly, but these effects are implicitly contained in its form as well; e.g., hydrophobic blocks will surely prefer to pack against each other. Although OPUS-PSP may be used in combination with other solvation models if necessary, it may be advantageous to avoid modeling explicit solvation effects in other cases. For example, in modeling membrane protein packing, OPUS-PSP may have an edge relative to other methods because the solvation dependence in this case may be very different from that of soluble proteins. Even though OPUS-PSP is constructed from a structure database of soluble proteins, the microenvironments of side-chain packing in membrane proteins should be similar to those of soluble proteins.

In constructing any statistical potential, the choice of reference state is very important.[36,37] The Boltzmann expression in eq 2 is a general way of developing the potential, and the accuracy of the potential can be improved by proper modeling of either p^{obs}, p^{ref}, or both. The significance of the choice of p^{ref} is evident in the development of the DFIRE[38] and DOPE[39] potentials. In OPUS-PSP, both p^{obs} and p^{ref} are modeled very differently, in which case the statistics of p^{obs} are generated based on the 19-block basis set and those of p^{ref} are generated by self-avoided random sampling of blocks with different sizes.[13] OPUS-PSP is also the first potential in which the geometry of interacting groups is explicitly considered in constructing the reference state.

OPUS-PSP is presently a discrete potential. In principle, it can be extended in two different ways. The first is to transform the discrete potential into a square-well potential and use it as a native contact potential between blocks. This is advantageous because the 19 blocks are expected to capture the essential elements of molecular interactions in an orientation-sensitive fashion. Such a contact potential can be combined with a funnel-like molecular mechanics potentials. In this way, OPUS-PSP may be used essentially as a bias to deepen the native state energy well without altering the long-range interactions. Note, the contact potential is short-range in nature, i.e., only sensitive to native-like packing patterns between blocks. The second is to revise OPUS-PSP to be continuous, so that derivatives can be obtained for molecular simulation.[40] However, a substantial reparameterization may be needed to achieve this.

A distinct feature of OPUS-PSP is that the interactions between pure main-chain atoms are excluded. However, many other studies showed that those interactions are important and highly correlated with the side-chain interactions.[4,5,41,42] Thus, revising the block basis set and including main-chain atoms may be directions for future improvement.

OPUS-PSP is a pairwise potential that allows for very rapid computational evaluation. This feature is critically important for some applications, such as the side-chain conformational modeling method, OPUS-Rota.[14] Along with its strong overall performance, OPUS-Rota performs particularly well in modeling aromatic side chains because of several design features. First, the contributions of aromatic residues in the rotamer frequency term are enhanced. Second, the vdW potential is softened for aromatic side chains, which enables the aromatic side chains to find their preferred rotamer angles, especially inside the densely packed protein core. Third, OPUS-PSP is inherently more sensitive to the orientation of the aromatic planes.

A major challenge in side-chain modeling is the issue of main-chain flexibility. The most successful methods, including OPUS-Rota, perform well when the main chain is in its native conformation, yet the accuracy of side-chain placement decreases quickly once the main chain deviates from its native state. There is of course a question of the significance of "native state" side-chain placement if the main chain is not in its native state. Main-chain and side-chain states are tightly coupled; if one is not in its native state, neither will the other. Thus, the ultimate way to solve this problem is to refine the main chain and side chain simultaneously.[43,44] There is another issue of causality between the main-chain and side-chain conformations. Most prediction methods try to position the main chain first and then place the side chains afterward. In reality, however, it is not unreasonable to assume that the main-chain conformation is dramatically influenced by side-chain packing. This is clear from the success of OPUS-PSP in decoy set recognition. OPUS-PSP does not explicitly account for pure main-chain interactions, yet it can consistently and accurately recognize the native state out of a large number of decoys. This result seems to imply that side-chain packing is crucial for native state formation; i.e., it is difficult to form a perfectly native protein backbone without having all of the side chains in place. This is also in line with the common observation that main-chain hydrogen-bonding interactions are not specific, because any pair of residues can form hydrogen bonds, while only specific pairs of side chains can be packed together favorably.

The author is thankful for the financial support of grants from the National Institutes of Health (R01-GM067801), the National Science Foundation (MCB-0818353), and the Welch Foundation (Q-1512). Critical reading of the manuscript by Athanasios D. Dousis and helpful discussion with Mingyang Lu are acknowledged.

BIOGRAPHICAL INFORMATION

Jianpeng Ma is a faculty member in the Department of Biochemistry and Molecular Biology at Baylor College of Medicine. He also holds a joint appointment in the Department of Bioengineering at Rice University. He received his undergraduate degree in physical chemistry from Fudan University in China. He received his doctoral degree in chemistry under the guidance of John E. Straub at Boston University, and then he moved to Harvard University for his postdoctoral training with Martin Karplus. He came to Houston in 2000 as an independent faculty member at Baylor and Rice. The research foci of Dr. Ma and his research group are on the development of novel multiscale computational methods for simulating, refining, and modeling flexible biomolecular complexes. Dr. Ma has received many awards for his contributions to biophysics and structural biology. The Welch Foundation awarded him the highly prestigious Norman Hackerman Award. He was elected as a Fellow of the American Physical Society (APS) and the American Association for the Advancement of Science (AAAS). He also won the Michael E. DeBakey Excellence in Research Award.

FOOTNOTES

* To whom correspondence should be addressed: One Baylor Plaza, BCM-125, Baylor College of Medicine, Houston, TX 77030. Telephone: 713-798-8187. Fax: 713-796-9438. E-mail: jpma@bcm.tmc.edu.

REFERENCES

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

2 Bahar, I.; Jernigan, R. L. Coordination geometry of nonbonded residues in globular proteins. *Fold. Des.* **1996**, *1*, 357–370.

3 Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997**, *18*, 849–873.

4 Buchete, N. V.; Straub, J. E.; Thirumalai, D. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* **2004**, *13*, 862–874.

5 Mukherjee, A.; Bhimalapuram, P.; Bagchi, B. Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.* **2005**, *123*, 014901.

6 Misura, K. M.; Morozov, A. V.; Baker, D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J. Mol. Biol.* **2004**, *342*, 651–664.

7 Miyazawa, S.; Jernigan, R. L. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins. *J. Chem. Phys.* **2005**, *122*, 024901.

8 Wu, Y.; Lu, M.; Chen, M.; Li, J.; Ma, J. OPUS-Ca: A knowledge-based potential function requiring only Cα positions. *Protein Sci.* **2007**, *16*, 1449–1463.

9 Buchete, N. V.; Straub, J. E.; Thirumalai, D. Dissecting contact potentials for proteins: Relative contributions of individual amino acids. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 119–130.

10 Buchete, N.-V.; Straub, J. E.; Thirumalai, D. Anisotropic coarse-grained statistical potentals improve the ability to identify native-like protein structures. *J. Chem. Phys.* **2003**, *118*, 7658–7671.

11 Makowski, M.; Sobolewski, E.; Czaplewski, C.; Liwo, A.; Oldziej, S.; No, J. H.; Scheraga, H. A. Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. 3. Calculation and parameterization of the potentials of mean force of pairs of identical hydrophobic side chains. *J. Phys. Chem. B* **2007**, *111*, 2925–2931.

12 Makowski, M.; Sobolewski, E.; Czaplewski, C.; Oldziej, S.; Liwo, A.; Scheraga, H. A. Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. IV. Pairs of different hydrophobic side chains. *J. Phys. Chem. B* **2008**, *112*, 11385–11395.

13 Lu, M.; Dousis, A.; Ma, J. OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* **2008**, *376*, 288–301.

14 Lu, M.; Dousis, A. D.; Ma, J. OPUS-Rota: A fast and accurate method for side-chain modeling. *Protein Sci.* **2008**, *17*, 1576–1585.

15 Samudrala, R.; Levitt, M. Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **2000**, *9*, 1399–1401.

16 Rajgaria, R.; McAllister, S. R.; Floudas, C. A. A novel high resolution Cα−Cα distance dependent force field based on a high quality decoy set. *Proteins* **2006**, *65*, 726–741.

17 Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **2003**, *53*, 76–87.

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

19 John, B.; Sali, A. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* **2003**, *31*, 3982–3992.

20 Gilis, D. Protein decoy sets for evaluating energy functions. *J. Biomol. Struct. Dyn.* **2004**, *21*, 725–736.

21 Lee, M. C.; Yang, R.; Duan, Y. Comparison between generalized-born and Poisson−Boltzmann methods in physics-based scoring functions for protein structure prediction. *J. Mol. Model.* **2005**, *12*, 101–110.

22 Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L., Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **2003**, *12*, 2001–2014.

23 Goldstein, R. F. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **1994**, *66*, 1335–1340.

24 Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **2002**, *356−542*, 539.

25 Xiang, Z.; Honig, B. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **2001**, *311*, 421–430.

26 Hartmann, C.; Antes, I.; Lengauer, T. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.* **2007**, *16*, 1294–1307.

27 Peterson, R. W.; Dutton, P. L.; Wand, A. J. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci.* **2004**, *13*, 735–751.

28 Liang, S.; Grishin, N. V. Side-chain modeling with an optimized scoring function. *Protein Sci.* **2002**, *11*, 322–331.

29 Dunbrack, R. L., Jr.; Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **1993**, *230*, 543–574.

30 Eisenberg, D.; McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199–203.

31 Sharp, K. A.; Nicholls, A.; Friedman, R.; Honig, B. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochemistry* **1991**, *30*, 9686–9697.

32 Zhang, N.; Zeng, C.; Wingreen, N. S. Fast accurate evaluation of protein solvent exposure. *Proteins* **2004**, *57*, 565–576.

33 Newman, M. E. J.; Barkema, G. T. *Monte Carlo Methods in Statistical Physics*; Clarendon Press: Oxford, U.K., 1999.

34 Jain, T.; Cerutti, D. S.; McCammon, J. A. Configurational-bias sampling technique for predicting side-chain conformations in proteins. *Protein Sci.* **2006**, *15*, 2029–2039.

35 Wallner, B.; Elofsson, A. All are not equal: A benchmark of different homology modeling programs. *Protein Sci.* **2005**, *14*, 1315–1327.

36 Betancourt, M. R.; Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **1999**, *8*, 361–369.

37 Chen, W. W.; Shakhnovich, E. I. Lessons from the design of a novel atomic potential for protein folding. *Protein Sci.* **2005**, *14*, 1741–1752.

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

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

40 Summa, C. M.; Levitt, M. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3177–3182.

41 Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16623–16633.

42 Fitzgerald, J. E.; Jha, A. K.; Colubri, A.; Sosnick, T. R.; Freed, K. F. Reduced Cβ statistical potentials can outperform all-atom potentials in decoy identification. *Protein Sci.* **2007**, *16*, 2123–2139.

43 Georgiev, I.; Donald, B. R. Dead-end elimination with backbone flexibility. *Bioinformatics* **2007**, *23*, I185−I194.

44 Li, G.; Liu, Z.; Guo, J.; Xu, Y. An algorithm for simultaneous backbone threading and side-chain packing. *Algorithmica* **2008**, *51*, 435–450.

45 Park, B.; Levitt, M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **1996**, *258*, 367–392.

46 Samudrala, R.; Xia, Y.; Levitt, M.; Huang, E. S. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.* **1999**, 505–516.

47 Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **2000**, *300*, 171–185.

48 Keasar, C.; Levitt, M. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* **2003**, *329*, 159–174.

49 Lin, M. S.; Fawzi, N. L.; Head-Gordon, T. Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure* **2007**, *15*, 727–740.

50 McConkey, B. J.; Sobolev, V.; Edelman, M. Discrimination of native protein structures using atom−atom contact scoring. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3215–3220.

51 Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* **2004**, *13*, 400–411.

52 Dehouck, Y.; Gilis, D.; Rooman, M. A new generation of statistical potentials for proteins. *Biophys. J.* **2006**, *90*, 4010–4017.

53 Dong, Q.; Wang, X.; Lin, L. Novel knowledge-based mean force potential at the profile level. *BMC Bioinf.* **2006**, *7*, 324.

54 Tobi, D.; Elber, R. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins* **2000**, *41*, 40–46.

55 Zhang, J.; Chen, R.; Liang, J. Empirical potential function for simplified protein models: Combining contact and local sequence-structure descriptors. *Proteins* **2006**, *63*, 949–960.

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

57 Colubri, A.; Jha, A. K.; Shen, M. Y.; Sali, A.; Berry, R. S.; Sosnick, T. R.; Freed, K. F. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *J. Mol. Biol.* **2006**, *363*, 835–857.

58 Wang, C.; Schueler-Furman, O.; Baker, D. Improved side-chain modeling for protein−protein docking. *Protein Sci.* **2005**, *14*, 1328–1339.

59 Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.