Published in final edited form as:

J Mol Biol. 2008 February 8; 376(1): 288–301. doi:10.1016/j.jmb.2007.11.033.

OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing

Mingyang Lu¹, Athanasios D. Dousis², and Jianpeng Ma^{1,2,*}

¹Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030

²Department of Bioengineering, Rice University, Houston, TX 77005

Abstract

In this paper, we report an orientation-dependent statistical all-atom potential derived from side-chain packing, named OPUS-PSP. It features a basis set of 19 rigid-body blocks extracted from the chemical structures of all 20 amino acid residues. The potential is generated from the orientation-specific packing statistics of pairs of those blocks in a non-redundant structural database. The purpose of such an approach is to capture the essential elements of orientation dependence in molecular packing interactions. Tests of OPUS-PSP on commonly used decoy sets demonstrate that it significantly outperforms most of the existing knowledge-based potentials in terms of both its ability to recognize native structures and consistency in achieving high Z-scores across decoy sets. As OPUS-PSP excludes interactions among main-chain atoms, its success highlights the crucial importance of side-chain packing in forming native protein structures. Moreover, OPUS-PSP does not explicitly include solvation terms, and thus the potential should perform well when the solvation effect is difficult to determine, such as in membrane proteins. Overall, OPUS-PSP is a generally applicable potential for protein structure modeling, especially for handling side-chain conformations — one of the most difficult steps in high-accuracy protein structure prediction and refinement.

Keywords

Knowledge-based Potential Function; Decoy Recognition; Structure Prediction; Protein Folding; Side-chain Packing; Orientation Dependence

Introduction

In the post-genomic era, despite major advances in high-throughput protein crystallization and other tools, modern experimental techniques in structural biology have not kept up with the rapid pace of genome sequencing. Consequently, the substantial sequence-structure gap is a driving force for research in protein structure prediction. However, accurately predicting the structure of a protein directly from its amino-acid sequence is far from routine due to various technical difficulties, one of which is the lack of accurate empirical potential functions.

*JM: To whom the correspondence should be addressed. Mailing address: One Baylor Plaza, BCM-125, Baylor College of Medicine, Houston, TX 77030. Email: jpmma@bcm.tmc.edu, Phone: 713-798-8187, Fax: 713-796-9438.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In general, there are two types of potential functions for protein structure modeling: physics-based and knowledge-based. Physics-based potential functions, such as the CHARMM force field¹, are developed from *ab initio* quantum chemical calculations, whereas knowledge-based potential functions are developed from statistical analysis of known protein structures. In many applications, knowledge-based potential functions outperform the physics-based potentials^{2; 3}. Knowledge-based potentials may be further divided into two categories: all-atom potentials and coarse-grained (semi)-residue-based potentials. For more than two decades, developing various potential functions has been an extremely active field. Many comprehensive review articles are available in the literature^{2; 4; 5; 6; 7; 8; 9; 10; 11; 12}.

One of the most important determinants of protein structure is side-chain packing, since the sequence identities of all polypeptide chains are solely designated by side-chains. In the literature, however, methods for generating statistical potentials are limited in their ability to describe side-chain packing. On the one hand, all-atom energy functions typically ignore the heterogeneous chemical bond connectivity¹³. For example, between two favorably packed side-chain groups, the proximity of individual atoms is substantially influenced by chemical bonding interactions. In other words, certain unfavorably interacting atoms can be near each other. Another issue is that, without proper description of chemical bonds, the orientational heterogeneity among packing groups is not well represented. On the other hand, typical (semi)-residue-based potentials insufficiently describe side-chain packing due to coarse-graining. The literature documents extensive efforts to exploit orientation preference for improving the description of side-chain packing^{14; 15; 16; 17; 18}, but much remains to be done to fully capture the orientation of side-chains when statistics are derived from known structures.

In this work, we have developed an orientation-dependent statistical potential based on side-chain packing (PSP), named OPUS-PSP. It is designed to overcome the aforementioned drawbacks and bridge the gap between all-atom and residue-based potentials. The foundation of OPUS-PSP hinges solely on side-chain packing interactions described by a unique basis set of rigid-body building blocks. This basis set is formed by decomposing the chemical structures of 20 amino acid residues into 19 block types. The potential is then generated from the orientation-specific packing statistics of pairs of those blocks in a non-redundant structural database. OPUS-PSP captures the essential details of side-chain flexibility and packing orientation without compromising the conciseness of a coarse-grained approach. Another feature of OPUS-PSP is that it does not consider mainchain-mainchain interactions, especially main-chain hydrogen-bonding interactions. The potential also does not explicitly include the solvation effect.

OPUS-PSP was tested on several commonly used decoy sets. The results demonstrate that it significantly outperforms most of the knowledge-based potentials in the literature, e.g., DFIRE¹⁹, DOPE²⁰/DOPE-C _{β} ⁴⁵, and HPMF²¹. Due to the exclusion of interactions among main-chain atoms, the success of the potential highlights the crucial importance of side-chain packing in building native protein structures. Furthermore, OPUS-PSP is widely applicable due to the generality of side-chain packing. In a forthcoming publication, we will show the effectiveness of the potential in improving side-chain modeling, one of the most important and difficult steps in high-accuracy protein structure prediction and refinement.

For clarity, the theoretical framework of OPUS-PSP is briefly outlined before the description of the main results, while the methodological details are given in the Methods section. Some lengthy technical issues that are of less importance to the general audience are explained in Supplemental Materials.

Theoretical Framework

In this section, we outline the theoretical framework of OPUS-PSP. Additional details of the implementation are given in the Methods section and in Supplemental Materials.

Two major components are involved in constructing OPUS-PSP: (a) the definition and geometry of 19 rigid-body blocks, and (b) a knowledge-based energy function derived from packing statistics of these blocks, combined with a repulsive Lennard-Jones term to deter steric clashes. We also apply coarse-graining and symmetry to improve the statistics.

Definitions of Rigid-body Blocks and Relative Orientation

To begin, we decompose the 20 residues into a structural basis of 19 rigid-body blocks. These blocks, shown in Fig. 1a, share three important characteristics: (a) all atoms in a block are bonded and in the same residue; (b) the bonded atoms are assumed to form a rigid body; (c) all heavy atoms are assumed to be in the same plane. Note that although assumptions (b) and (c) are invalid for the proline ring of block type 19, these assumptions are found to be reasonable for our potential. Furthermore, we ignore the alpha carbon atoms of all residues except Pro and Gly because we assume that shielding by bonded atoms around alpha carbons reduces the influence of those alpha carbons on side-chain packing (our results support this assumption). Each residue contains more than one block, but each block appears only once in a single residue. The block compositions of the 20 residue types are summarized in Fig. 1b. 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 .

OPUS-PSP requires a coordinate system to define the relative orientation of a pair of blocks. In order to make it permutationally invariant (i.e., order of blocks does not matter), we express the relative orientation of block types a and b using three variables: two relative direction vectors $\mathbf{r}_{a \rightarrow b}$ and $\mathbf{r}_{b \rightarrow a}$, and an inter-rotation angle ψ_{ab} along the axis connecting the origins of the two blocks in their respective molecular reference frames, as illustrated in Fig. 2. For a pair of blocks, these three variables (computed in the laboratory reference frame), coupled with the molecular reference frame for each block, completely define the relative orientation of that pair of blocks, i.e., the axial rotation around the line linking the origins of the two blocks and the pivot motion around the origin of each block. The origin and molecular reference frame for each block type and the inter-rotation angle are specified in Supplemental Materials.

Energy Function

OPUS-PSP is composed of an orientation-dependent packing energy term E_{orient} and a repulsive energy term E_{repl} :

$$E_{\text{PSP}} = E_{\text{orient}} + w_{\text{repl}} E_{\text{repl}} \quad (1)$$

where w_{repl} is a weight parameter optimized against a small subset of decoy sets (see Methods).

To calculate the total orientation-dependent packing energy, E_{orient} , we first define the packing energy for a pair of blocks in relative orientation space using the Boltzmann formula:

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

Here, p^{obs} is the probability of a particular orientation state for block types a and b in contact (“contact probability”; see Methods for the definition of block contact) with respect to all

observed contact states for any block pair extracted from the non-redundant structure database, p^{ref} is the contact probability of all possible occurrences of that state without packing interactions (the reference state), $\Omega_{ab} = (\mathbf{r}_{a \rightarrow b}, \mathbf{r}_{b \rightarrow a}, \psi_{ab})$ specifies the relative orientation of a and b , and $k_B T$ is the Boltzmann constant (set to unity). Thus, to derive the packing energy function, we must: (a) count the number of native contacts for all relative orientations of all pairs of blocks using a non-redundant protein structure database (p^{obs}); and (b) estimate the reference state that encapsulates all possible packing orientations (p^{ref}). These two steps are explained in detail in Methods.

The total orientation-dependent packing energy is calculated by summing the packing energies of all pairs of blocks in contact (“block contact pairs”) between all pairs of non-consecutive residues:

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

Here, $\delta(\alpha, \beta)$ is one when blocks α and β are in contact and zero otherwise, $B(\alpha) = a$ maps block α to its block type a , and $\widehat{E}(a, b) = n(a, b)E(\Omega_{ab}, a, b)$, where $n(a, b)$ is the average number of pairs of heavy atoms in contact (“atom contact pairs”; two atoms are in contact when their pairwise distance is less than 5 Å) between block types a and b , used as a weighting term for block size. The weighting term is calculated by random sampling in the manner of the reference state probability calculation (see Methods), and it is necessary because larger blocks contribute more atom contact pairs and therefore more energy. In calculating E_{orient} , we restrict our calculation to sidechain-sidechain and mainchain-sidechain interactions only. The mainchain-mainchain hydrogen bonding and other short-range interactions are not considered. Thus, we ignore all contact pairs of block types {1, 5, 6, 7}, as types 1 and 7 are main-chain functional groups and types 5 and 6 have contact interactions that are highly correlated to those of the main-chain atoms. The sensitivity of E_{orient} to the features above is analyzed in the Results section.

The repulsive term E_{repl} is necessary because a purely orientation-dependent packing function cannot deter steric clashes, and it is defined as:

$$E_{\text{repl}} = \sum_{i, j} E_{\text{LJ}}(i, j) \quad (4)$$

where $E_{\text{LJ}}(i, j)$ is a repulsive Lennard-Jones (LJ) potential (no attractive term) for two atoms i and j . Like E_{orient} , the summation in the LJ term excludes interactions between pairs of main-chain atoms and between two atoms in the same residue. Note that OPUS-PSP minimizes over-counting of similar interactions because E_{orient} and E_{repl} are typically orthogonal.

Coarse-graining of Orientation Bins and Symmetry

Given the limited amount of non-homologous protein data available for building the energy function, it is necessary to coarse-grain the orientation space and exploit the symmetry of the 19 blocks. These blocks are classified into nine symmetry classes that belong to two basic groups: plane shapes (IV, V, VII-IX) and line shapes (I-III, VI), as shown in Fig. 1a. Note that VI is considered a line shape because of the six-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. These bins are illustrated in Fig. 3a. For each line-shaped block, the cylindrical symmetry permits usage of five latitudinal bins as shown in

Fig.3b. Fig.3c illustrates 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.

Thus, for two blocks in contact, the maximal number of bins is $26 \times 4 \times 26 = 2704$. However, in practice, we combine certain redundant bins based on the intrinsic molecular symmetry of the blocks, such that the effective number of bins is much smaller. For details of this symmetry treatment, see Supplemental Materials.

To balance the trade-off between the number of bins and the available structure data for statistical analysis, we find that a choice of 26 directional bins is appropriate for plane-shaped blocks. In a coarser approximation, we tried six bins ($\pm\mathbf{x}, \pm\mathbf{y}, \pm\mathbf{z}$) and the results were much worse. Using more than 26 bins would severely deplete the structure data.

Results

Decoy Set Recognition

The predictive power of OPUS-PSP was tested in benchmark studies using the popular decoy set collections: Decoys 'R' Us²², HR²³, Rosetta (and Rosetta2)^{24; 25}, MOULDER²⁶, structural (<http://dd.compbio.washington.edu/>), and the decoy sets collected by Gilis²⁷, which we call the Gilis collection. The results are summarized in Table 1.

First, we compare our potential to twelve other energy functions using the 34 decoy sets of Decoys 'R' Us. The native conformation rankings and Z-scores for all decoy sets are listed in Table 1b. OPUS-PSP successfully identifies the native conformation for 31 of the 34, surpassing all twelve potentials in this benchmark. In particular, OPUS-PSP recognizes the native conformation of 1b0n-B, whereas MSE²⁸, DOPE²⁰, and HPMF²¹ cannot. Another important feature is that the Z-scores of OPUS-PSP are very consistent for all the successful decoy sets, indicating the robustness of the potential. It is worth noting that, in sharp contrast to the 31 successful cases, the rankings of the native structures in the three decoy sets that OPUS-PSP failed to recognize (1fc2 in fisa, 1bba in lmds, and 1fc2 in lmds) are far from the top. Most of the other potentials perform unsatisfactorily for these decoy sets as well. We believe that this is mainly due to the intrinsic nature of the decoy sets, e.g., having poorly packed native structures, rather than to the predictive power of the potentials themselves.

OPUS-PSP's performance on the MOULDER decoy set collection is similar to that of DOPE²⁰, Rosetta^{16; 25; 29}, DFIRE¹⁹, and DFIRE-SCM³⁰. The single structure that all of the tested energy functions fail to correctly identify is an NMR structure (PDB code: 2pna).

On the HR decoy set collection, which contains many high-resolution near-native decoys, OPUS-PSP performs very well. It recognizes 135 of 148 native conformations, and of the thirteen missed, six are still ranked within the top five.

On the Rosetta and Rosetta2 decoy set collections, OPUS-PSP performs very well on structures solved by X-ray crystallography, whereas it is less successful on the broader set that includes structures solved by NMR spectroscopy²⁰. This is reasonable given that our potential is trained using X-ray structures. OPUS-PSP considerably outperforms DFIRE¹⁹, DFIRE-SCM³⁰, and CALSP³¹ on the X-ray structure subset of the Rosetta decoys. For the more challenging Rosetta2 collection, which contains a large number of decoys that are biased toward the native structure, OPUS-PSP recognizes more than twice the native conformations that DOPE recognizes.

Out of all the benchmarks, OPUS-PSP is outperformed only on the structural decoy sets by the MM-PBSA³² and MJ_2005 potentials¹⁷. These decoy sets contain decoys generated by

comparative modeling of globins and immunoglobulins, ~ 60% of which have a C α RMSD (from the native conformation) less than 2.5 Å. For the ig_structal and ig_structal_hires sets, we find that OPUS-PSP performs better if main-chain interactions between pairs of block types {1,5,6,7} are also included in the total energy calculation (see Theoretical Framework).

The potential with the closest performance to OPUS-PSP is the very recent hydrophobic potential of mean force (HPMF) model²¹. OPUS-PSP outperforms HPMF in the number of native conformations identified and in mean Z-score for both Decoys 'R' Us and the combined set of X-ray structures from the Rosetta and Rosetta2 collections ("Rosetta1+2" in Table 1a)²¹. Moreover, HPMF contains many energy parameters that are trained from a subset of decoys, whereas OPUS-PSP has only one adjustable weight for two nearly independent terms, and thus OPUS-PSP's performance is less dependent upon the choice of training set.

Sensitivity Analysis of Energy Function Features

We conducted a sensitivity analysis to identify the features of OPUS-PSP that are most important for correctly identifying native conformations within a set of decoys. First, we tested the orientation-dependent packing energy term E_{orient} of OPUS-PSP in three aspects: (a) orientation preference, (b) weighting by block size, and (c) main-chain interactions among blocks types {1,5,6,7}. We generated energy functions for all combinations of these three aspects, and we tested the resulting eight energy functions (including the energy function described by Equ.3) on the 34 decoy sets in Decoys 'R' Us.

To test the importance of orientation preference for two block types a and b that are in contact, we compared E_{orient} to a simple contact energy function with no preference for orientation:

$$E(a,b) = -k_B T \log \frac{p^{\text{obs}}(a,b)}{p^{\text{ref}}(a,b)}.$$

Here, $p^{\text{obs}}(a,b) = N^{\text{obs}}(a,b) / \sum_{a,b} N^{\text{obs}}(a,b)$ and $p^{\text{ref}}(a,b) \approx \chi(a)\chi(b)$, where

$\chi(a) = N^{\text{obs}}(a) / \sum_a N^{\text{obs}}(a)$ is the mole fraction of block type a . Results listed in Table 2a indicate the overwhelming importance of the orientation preference.

To test the effect of weighting the pairwise packing energies by block size, we compared the decoy recognition results with and without the size-weighting term $n(a,b)$ from Equ.3. Results listed in Table 2a indicate that the weighting has a small effect on performance.

To examine the relevance of main-chain interactions among block types {1,5,6,7}, we tested E_{orient} with and without the pairs of {1,5,6,7}. Block types 1 and 7 are obvious since they are functional groups from the main-chain, and we also include 5 and 6 because they typically fall very close to the main-chain and are therefore expected to be highly correlated with main-chain interactions. Results listed in Table 2a indicate that the exclusion of those interactions slightly improves decoy set recognition.

In summary for E_{orient} , orientation preference is most important, improving native conformation recognition more than two-fold. E_{orient} is least sensitive to weighting by block size, but we keep the weighting term because it is natural to emphasize the packing energy of big blocks.

Finally, we also examined the significance of the repulsive term E_{repul} in OPUS-PSP by testing all decoy sets with and without this term. The results are shown in Table 2b. With E_{repul} , the performance of OPUS-PSP increases for all decoy sets except for the HR and structal decoy

sets. In the case of HR, OPUS-PSP without E_{repul} recognizes seven more native conformations than with E_{repul} , but of those seven missed with E_{repul} , six native conformations are still ranked among the top five. In the ig_structal_hires set, OPUS-PSP without E_{repul} outperforms all other potentials, recognizing all 20 native conformations.

Most-favored Interaction Patterns

In our study, we found that the packing between certain block types present clear patterns consistent with a previous analysis of anisotropic side-chain packing¹⁶. In Table 3 and Fig.4, we show the 30 most-probable and unique packing patterns extracted from the non-redundant structure database. They are presented in terms of their chemical nature of interaction, e.g., hydrophobic, electrostatic, π - π and cation- π interactions.

Several universal patterns may be identified from these examples. For instance, aromatic side-chains tend to form certain patterns of π - π and cation- π interactions. The packing patterns for hydrophobic interactions favor contact at the branching points more than at the tips of the branches, as in cases 12-12ii, 12-15, and 15-15. The presence of nitrogen and oxygen atoms also noticeably influences the packing patterns.

Discussion

In this paper, we report an orientation-dependent statistical all-atom potential. Amino acid residues are constructed from a basis set of 19 rigid-body blocks, and the potential is then established by statistical analysis of the orientation-specific packing patterns among those blocks in a non-redundant structural database. The purpose of this scheme is to capture the essential details of side-chain flexibility and packing orientation. Tests of OPUS-PSP on commonly used decoy sets show that it significantly outperforms most of the knowledge-based potentials in the literature, especially the recent DFIRE¹⁹, DOPE²⁰/DOPE- C_{β} ⁴⁵, and HPMF²¹ potentials. OPUS-PSP is expected to be particularly useful in modeling side-chain conformations, one of the most important and difficult steps in high-accuracy protein structure prediction and refinement.

OPUS-PSP has many advantages in terms of its coarse-graining scheme, interaction parameters and reference state. In terms of OPUS-PSP's coarse-graining scheme, the unique 19 rigid-body basis set captures the essential elements of orientation dependence in molecular packing interactions. For example, OPUS-PSP can sense subtle changes in even a single rotamer conformation because the orientations of the constituent blocks are significantly changed as a consequence.

In terms of the orientational parameters, a key trade-off exists between the number of parameters and their sensitivity in describing the orientation. Choosing too many parameters risks overfitting the limited data, yet the packing pattern sensitivity of the potential suffers if the orientational parameters are too coarse. To achieve a better balance, we extensively use block symmetry to consolidate redundant bins. We also employ Bayesian statistics to compensate for the cases in which the available data are sparse.

The reference state is also important in constructing a statistical potential. In OPUS-PSP, the reference state is calculated by random sampling of blocks, so the excluded-volume effects and block-specific geometries are naturally taken into account. To the best of our knowledge, this is the first case that the geometry of interacting groups is explicitly considered in constructing the reference state.

We want to emphasize that OPUS-PSP is built almost exclusively on the orientation-specific packing interactions of side-chains. Only a repulsive energy term is added to prevent steric

clash. In fact, we found that the potential without the repulsive term performs very well by itself in decoy set recognition. The repulsive term is required for the more general purpose of structure prediction.

No main-chain hydrogen bonding interactions are explicitly considered in OPUS-PSP. One reason is that the rigid-body basis set used for constructing OPUS-PSP is not suited for optimizing main-chain hydrogen bonding interactions. Another reason is that main-chain hydrogen bonding interactions alone are not very specific in terms of pairing between two residues, i.e., any pair of residues can form main-chain hydrogen bonds, while not all pairs of side-chains can pack against each other favorably. As clearly demonstrated in our results, OPUS-PSP, which excludes the interactions among main-chain atoms, performs very well.

It is interesting to note that a very recent paper by Fitzgerald et al.⁴⁵ indicates that a main-chain potential constructed from the DOPE potential (with each side-chain reduced to a single C β atom) can do better than the all-atom DOPE potential in decoy set recognition. This is an opposite extreme of our approach. Although both approaches are successful in decoy set recognition, we believe that they will have their own distinct domains of application. For example, OPUS-PSP will be highly advantageous in modeling side-chain rotamer conformations.

In addition, no explicit solvation terms are included in OPUS-PSP. Although the side-chain packing interactions are not completely orthogonal to the solvation effect, e.g., hydrophobic side-chains prefer to pack together, it is advantageous to explicitly avoid the solvation effect if doing so does not impair performance. One reason is that OPUS-PSP is statistically formulated from a structure database of soluble proteins. If the solvation effect is explicitly included, the potential would be less reliable for systems such as membrane proteins with a very different solvation environment. We expect the current form of OPUS-PSP to perform well in modeling membrane proteins, as their side-chain packing interactions should be very similar to those of soluble proteins.

Methods

As already outlined in the Theoretical Framework section, the packing energy function requires: (a) the proportions of observed native block-pair contacts for all relative orientations; and (b) the reference state that encapsulates the idealized preference of all possible packing orientations. In addition, we apply Bayesian inference to estimate the statistics of block-pair-orientations that occur infrequently in the protein structure database.

Counting Pair Contacts from a Non-homologous Structure Database

To determine p^{obs} , the observed fraction of occurrences of a particular pair-orientation state with respect to all observed states (see Equ.2), we tally the number of contacts N^{obs} between all pairs of blocks at all relative orientations that exist within a representative, non-homologous structure database, i.e., $p^{\text{obs}}(\Omega_{ab}, a, b) = N^{\text{obs}}(\Omega_{ab}, a, b) / N_{\text{total}}^{\text{obs}}$, and $N_{\text{total}}^{\text{obs}} = \sum_{a,b} N^{\text{obs}}(\Omega_{ab}, a, b)$. Contact occurs when any heavy atom from one block falls within a 5Å radius of any heavy atom from another block. Our non-homologous structure database of 3211 protein chains is generated from the PISCES server³³ using a percent identity cutoff of 40%, a resolution cutoff of 1.8Å, and an R-factor cutoff of 0.25. The 40% percent identity cutoff is a looser criterion than those of most other knowledge-based potentials in the literature, which increases the available set of structures and thereby improves the statistics. We argue that side-chain packing preference is less sensitive to the percent identity. In addition, our test results on decoy set recognition are very similar when comparing our current version and a version built from the structure database with 30% percent identity cutoff.

Calculating the Reference State Probability

The reference state probability $p^{\text{ref}}(\Omega_{ab}, a, b)$ in Equ.2 is calculated as the product of the contact-pair-orientation statistic $p^{\text{ref}}(\Omega_{ab} | a, b)$ and the contact probability $p^{\text{ref}}(a, b)$:

$$p^{\text{ref}}(\Omega_{ab}, a, b) = p^{\text{ref}}(\Omega_{ab} | a, b) p^{\text{ref}}(a, b). \quad (5)$$

We use random sampling to compute $p^{\text{ref}}(\Omega_{ab} | a, b)$, since no obvious analytical solution exists. Two blocks are randomly placed in space without consideration for chemical interactions.

Block pairs are discarded if they are out of contact range ($>5\text{\AA}$) or clashing ($<2 \times 2.0 / \sqrt[6]{2}\text{\AA}$, where the atomic radius of a carbon atom is $2.0\text{\AA}$). Otherwise, the relative orientation of the block pair is determined and tabulated. A large number of such contact pairs (an average of 10^5 counts/bin) are generated to minimize the sampling error.

The calculation of the block contact probability $p^{\text{ref}}(a, b)$ is not straightforward, in contrast to the estimation of contact probability for two residues or two atoms, which is intensively discussed in the literature^{13; 19; 34; 35}. Typically, residues are assumed to be of similar size, and thus the quasichemical approximation based on mole fractions of the residues is reasonable³⁴. We use this approximation as the starting point for estimating the contact probability for two blocks. First, let the mole fraction of block type a be

$$\chi(a) = \sum_m \chi(m) \delta(a, m),$$

where $\chi(m) = N^{\text{obs}}(m) / N_{\text{total}}^{\text{obs}}$ is the mole fraction of residue type m , $\delta(a, m)$ is one if a is in m and zero otherwise, and the summation is over all 20 residue types. To represent the connectivity of the blocks within a residue, we assume that each block type a in a residue type m has an equal and constant probability ξ of being in contact with any block type b of residue type n if m and n are in contact. Hence, from the quasichemical approximation

$$p^{\text{ref}}(a, b) \approx \sum_{m, n} \chi(m) \chi(n) (\xi \cdot \delta(a, m) \delta(b, n)) = \xi \chi(a) \chi(b).$$

Next, since $\sum_{a, b} p^{\text{ref}}(a, b) = \sum_{a, b} \chi(a) \chi(b) = 1$, then ξ must be unity, and

$$p^{\text{ref}}(a, b) \approx \chi(a) \chi(b). \quad (6)$$

Equ.6 implies that the quasichemical approximation is appropriate for estimating the contact probabilities of two blocks. Note that the equation above does not explicitly consider the size differences of residues and blocks. Presumably, larger blocks or residues have more contacts than smaller ones. However, the energy functions derived from reference states that incorporate contact number are found to be less powerful than those that do not³⁴. Our current study also supports this conclusion, as our energy function using the quasichemical approximation so far produces the best results.

Rare Data Estimation

Given the relatively small size (3211 chains) of the non-homologous protein dataset used to build our potential, many pair-orientations occur too infrequently for the data to be statistically

significant. Thus, we use Bayesian statistics to solve this rare data problem, the details of which are described in Supplemental Materials.

The Repulsive Lennard-Jones (LJ) Potential Term

Similar to that used in the literature³⁶, the repulsive LJ potential function for two atoms i and j takes the form

$$E_{LJ}(i,j) = \begin{cases} 49.69 - 40.06d_{ij}^*, & d_{ij}^* \in [0, 1/1.33] \\ e_{ij} \left[(d_{ij}^*)^{-12} - 2(d_{ij}^*)^{-6} \right] & d_{ij}^* \in (1/1.33, 1/1.12] \\ 0, & d_{ij}^* \in (1/1.12, \infty) \end{cases} \quad (7)$$

where $d_{ij}^* = d_{ij}/a_{ij}$, d_{ij} is the distance between atoms i and j , $a_{ij} = a_i + a_j$, $e_{ij} = \sqrt{e_i e_j}$, and a_i , a_j and e_i , e_j are atomic radii and well depths for the LJ calculations, respectively. Thus, $E_{LJ}(i, j)$ is linear in the region of severe steric clash, and it is zero in (i.e., excludes) the attractive region of the LJ potential. In order to allow hydrogen and disulfide bonds, a_{ij} is specifically chosen for atoms involving hydrogen or disulfide bonds. The detailed LJ parameters are described in Supplemental Materials.

The total LJ energy is calculated by summing over all possible pairs of heavy atoms except in the following three cases: (a) pairs of main-chain heavy atoms including C β , as its position is largely determined by the main-chain; (b) two atoms in the same residue; and (c) Pro C δ and C of the previous residue, as they are separated by only two bonds.

Weight Optimization for the Combined Energy Function

To determine the optimal weight parameter w_{repul} , we run multiple decoy recognition trials using w_{repul} ranging from 0.0 to 2.0 at 0.1 intervals. The training set includes all 7 decoy sets in the 4state_reduced subset of Decoys 'R' Us and the 13 decoy sets in the Gilis collection that are not included in other benchmarks in this work. The decoys in 4state_reduced have fewer steric clashes than the decoys in the Gillis collection. Thus, we can use both collections in the optimization to find a balance between orientation-dependence and steric-clash-avoidance. From our results, $w_{\text{repul}} = 1.2$ is the optimal weight, as it produces the best native conformation rankings and Z-scores. In any case, the decoy recognition performance of OPUS-PSP is not sensitive to the weight value.

As an added benefit, this weight optimization is fairly straightforward with few disadvantages. First, the risk of overfitting is reduced because there is only one weight parameter to optimize. Second, E_{orient} and E_{repul} are nearly independent, such that the combined energy is unlikely to suffer from over-counting. Third, our training set contains decoys generated from various methods, thereby balancing the contributions of the energy terms. This weight optimization method, which is very similar to that of HPMF, is supported by our tests of OPUS-PSP on the other decoy sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

JM acknowledges financial support by a grant from the National Institutes of Health (R01-GM067801) and a grant from the National Science Foundation CAREER Award (MCB-0237796). ML is partially supported by a pre-doctoral fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia. AD is

supported by a pre-doctoral fellowship from the Keck Center through the National Library of Medicine Computational Biology and Medicine Training Program (NLM Grant No. 5T15LM07093).

References

- MacKerell AD, Bashford D Jr, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher I, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem* 1998;B102:3586–3616.
- Skolnick J. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* 2006;16:166–71. [PubMed: 16524716]
- Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. *Proteins* 2005;61(Suppl 7):128–34. [PubMed: 16187354]
- Sippl MJ. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 1995;5:229–35. [PubMed: 7648326]
- Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* 1996;6:195–209. [PubMed: 8728652]
- Moult J. Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol* 1997;7:194–9. [PubMed: 9094335]
- Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* 2000;10:139–145. [PubMed: 10753811]
- Gohlke H, Klebe G. Statistical potentials and scoring functions applied to protein-ligand binding. *Curr Opin Struct Biol* 2001;11:231–5. [PubMed: 11297933]
- Russ WP, Ranganathan R. Knowledge-based potential functions in protein design. *Curr Opin Struct Biol* 2002;12:447–52. [PubMed: 12163066]
- Buchete NV, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. *Curr Opin Struct Biol* 2004;14:225–32. [PubMed: 15093838]
- Poole AM, Ranganathan R. Knowledge-based potentials in protein design. *Curr Opin Struct Biol* 2006;16:508–13. [PubMed: 16843652]
- Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? *Cell Biochem Biophys* 2006;46:165–74. [PubMed: 17012757]
- Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. *Protein Sci* 2005;14:1741–52. [PubMed: 15987903]
- Buchete NV, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci* 2004;13:862–74. [PubMed: 15044723]
- Mukherjee A, Bhimalapuram P, Bagchi B. Orientation-dependent potential of mean force for protein folding. *Journal of Chemical Physics* 2005;123
- Misura KM, Morozov AV, Baker D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J Mol Biol* 2004;342:651–64. [PubMed: 15327962]
- Miyazawa S, Jernigan RL. 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. [PubMed: 15638624]
- Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: A Knowledge-based Potential Function Requiring Only C α Positions. *Prot. Sci* 2007;16:1449–1463.
- 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;11:2714–26. [PubMed: 12381853]
- Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* 2006;15:2507–24. [PubMed: 17075131]
- Lin MS, Fawzi NL, Head-Gordon T. Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure* 2007;15:727–40. [PubMed: 17562319]

22. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci* 2000;9:1399–401. [PubMed: 10933507]
23. Rajgaria R, McAllister SR, Floudas CA. A novel high resolution Calpha--Calpha distance dependent force field based on a high quality decoy set. *Proteins* 2006;65:726–41. [PubMed: 16981202]
24. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* 2003;53:76–87. [PubMed: 12945051]
25. Simons KT, 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;268:209–25. [PubMed: 9149153]
26. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res* 2003;31:3982–92. [PubMed: 12853614]
27. Gilis D. Protein decoy sets for evaluating energy functions. *J Biomol Struct Dyn* 2004;21:725–36. [PubMed: 15106995]
28. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. *Proc Natl Acad Sci U S A* 2003;100:3215–20. [PubMed: 12631702]
29. Simons KT, Ruczinski I, Kooperberg C, Fox BA, 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. [PubMed: 10336385]
30. 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;13:400–11. [PubMed: 14739325]
31. Zhang J, Chen R, Liang J. Empirical potential function for simplified protein models: combining contact and local sequence-structure descriptors. *Proteins* 2006;63:949–60. [PubMed: 16477624]
32. Lee MC, 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;12:101–10. [PubMed: 16096807]
33. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–91. [PubMed: 12912846]
34. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci* 1997;6:676–88. [PubMed: 9070450]
35. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 1996;256:623–44. [PubMed: 8604144]
36. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science* 2003;302:1364–8. [PubMed: 14631033]
37. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol* 1996;258:367–92. [PubMed: 8627632]
38. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac Symp Biocomput* 1999:505–16. [PubMed: 10380223]
39. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol* 2000;300:171–85. [PubMed: 10864507]
40. 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;329:159–74. [PubMed: 12742025]
41. Dehouck Y, Gilis D, Rooman M. A new generation of statistical potentials for proteins. *Biophys J* 2006;90:4010–7. [PubMed: 16533849]
42. Dong Q, Wang X, Lin L. Novel knowledge-based mean force potential at the profile level. *BMC Bioinformatics* 2006;7:324. [PubMed: 16803615]
43. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins* 2000;41:40–6. [PubMed: 10944392]

44. Colubri A, Jha AK, Shen MY, Sali A, Berry RS, Sosnick TR, Freed KF. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *J Mol Biol* 2006;363:835–57. [PubMed: 16982067]
45. Fitzgerald JE, Jha AK, Colubri A, Sosnick TR, Freed KF. Reduced C_{β} statistical potentials can outperform all-atom potentials in decoy identification. *Protein Sci* 2007;16:2123–39. [PubMed: 17893359]

Figure 1.

Definition of 19 rigid-body blocks. (a) Illustration of block types. Blocks are categorized into nine symmetry classes from I to IX. Block classes I, II, III and VI are line shapes, and the others are plane shapes. Heavy atoms are numbered for determining contacts, steric clashes and reference frames (see Supplemental Materials). R and R' are not considered parts of the blocks but are shown to indicate connectivity only. The reference frames for line shapes and plane shapes are schematically shown alongside their corresponding block types at the bottom of the figure. The origins of the reference frames are explicitly defined in Supplemental Materials. (b) Block composition of residues. All blocks are circled for all amino acids. The block types

(numbers in parentheses) are defined in (a). Each residue contains at least one block type, and within that residue, no block type occurs more than once.

Figure 2. The definition of relative orientation. If block types a and b are in contact, then $\mathbf{r}_{a \rightarrow b}$ and $\mathbf{r}_{b \rightarrow a}$ 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.

Figure 3.

The definition of the relative direction bins for line-shaped and plane-shaped blocks. (a) Relative direction bins for plane-shaped blocks (classes IV, V, VII, VIII and IX). The complete set of 26 bins denoted by the index $(n_x n_y n_z)$ is derived from the spherical angles θ and ϕ of vector $\mathbf{r}_{a \rightarrow b}$ in the reference frame of block a . (b) Relative direction bins for line-shaped blocks (classes I, II, III and VI). The complete set of 5 bins denoted by the index $(n_x n_y)$ is derived from the angle θ between the primary axis (x-axis) and vector $\mathbf{r}_{a \rightarrow b}$ formed from the origin $\mathbf{o}_a$ of block a to the origin $\mathbf{o}_b$ of block b . (c) The 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 bins for line shapes (on the right) are consolidated from the latitudinal bins of plane shapes.

Figure 4.

Examples of most-probable packing patterns (30 in total). For each pattern, the numbers represent the indices of the block types in contact, and the lowercase roman numerals specify different contact patterns for a given pair of block types. For most block pairs, the consistent clustering of contact patterns provides partial validation of OPUS-PSP. Note that examples such as 17-17i and 12-12i have multiple clusters due to symmetry. Table 3 briefly summarizes the contact patterns illustrated here.

Table 1**OPUS-PSP performance on various decoy sets**

(a) OPUS-PSP performance compared to other potentials (b) OPUS-PSP performance on Decoys 'R' Us

	(a)	
	Top 1/Total No. ^a	Mean
Decoys 'R' Us ^{25; 37; 38; 39; 40}		
<i>OPUS-PSP</i>	31/34	-5.37
HPMF ²¹	29/32 ^b	-4.18
DOPE ²⁰	28/32	--
MSE ²⁸	21/23	-5.78
DFIRE ¹⁹	27/32	-4.52
MJ_2005 ¹⁷	27/34	-5.93
DFIRE-SCM ³⁰	23/32	-4.36
MM-PBSA ³²	23/34	-1.95
DGR ⁴¹	21/25	-5.25
DWL ⁴²	21/32	-3.66
TE13 ⁴³	14/25	-3.53
CALSP ³¹	15/25	--
Rosetta ^{16; 25; 29}	14/32 ^c	--
MOULDER ²⁶		
<i>OPUS-PSP</i>	19/20	-4.60
DOPE	19/20 ^c	--
Rosetta	19/20 ^c	--
DFIRE	19/20 ^c	--
DFIRE-SCM	19/20 ^c	--
HR ²³		
<i>OPUS-PSP</i>	135/148	-7.50
HR ²³	113/150	--
TE13	92/148 ^d	--
Rosetta (X-ray) ²⁵		
<i>OPUS-PSP</i>	37/41	-6.56
DFIRE	31/41	-3.91
DFIRE-SCM	33/41	-4.90
CALSP	28/41	-4.16
Rosetta2 ^{24; 25}		
<i>OPUS-PSP</i>	23/41	-2.71
<i>OPUS-PSP</i> (X-ray)	22/25	-4.49
DOPE	11/41 ^e	-1.50

(a)		
	Top 1/Total No. ^a	Mean
Rosetta 1+2^f (X-ray)^{24; 25}		
OPUS-PSP	34/35	-6.76
HPMF	30/35	-4.42
hg_structal^g		
OPUS-PSP	18/29	-1.76
MM-PBSA	20/29	-1.60
MJ_2005	22/29	-2.76
ig_structal^g		
OPUS-PSP	46/61 ^h	-2.79
MJ_2005	49/61	-3.55
ig_structal_hires^g		
OPUS-PSP	19/20 ^h	-3.03
MJ_2005	19/20	-4.31
Gilis²⁷		
OPUS-PSP	43/45	-5.58

(b)				
	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
fisa_casp3				
12	1bg8-A	1201	1	-6.01
13	1bl0	972	1	-6.00

(b)				
	PDB code	Decoy set size	Rank	Z-score
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“Total No.” 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.

^bOPUS-PSP recognizes 30 of the 32 decoy sets used for HPMF

^cResults taken from ²⁰

^dResults taken from ²³

^eResults taken from ⁴⁴

^fThe total number of 35 is a subset of X-ray structures in the combined Rosetta and Rosetta2 collections

^gFrom <http://dd.compbio.washington.edu/>

^hOPUS-PSP includes main-chain interactions of block types {1,5,6,7}.

Table 2

Sensitivity analysis of OPUS-PSP features

(a) Sensitivity analysis of E_{orient} features using Decoy 'R' Us. The notation "+/-" denotes inclusion/exclusion of a particular feature. The column labeled "{1,5,6,7}" represents the interactions among block types {1,5,6,7}. The integer values are the number of native conformations ranked first by OPUS-PSP (i.e., correctly recognized), and the mean Z-scores are in parentheses. (b) The comparison of OPUS-PSP with and without the repulsive term E_{repul} using various decoy sets. Note that the version of OPUS-PSP with E_{repul} is used for Table 1

Orientation Preference			
Weight	{1,5,6,7}	+	-
+	-	30 (-5.36)	12 (-2.43)
-	-	30 (-5.31)	13 (-2.46)
+	+	24 (-4.63)	14 (-2.27)
-	+	24 (-4.14)	12 (-1.81)

Repulsive Term		
Decoy Sets	Without	With
Decoy 'R' Us	30/34	31/34 (-5.37)
Rosetta	45/92	55/92 (-4.17)
Rosetta (X-ray)	33/41	37/41 (-6.56)
Rosetta2	17/41	23/41 (-2.71)
Rosetta2 (X-ray)	16/25	22/25 (-4.49)
Rosetta 1+2 (X-ray)	30/35	34/35 (-6.76)
MOULDER	19/20	19/20 (-4.60)
HR	142/148	135/148 (-7.50)
hg_structural	18/29	18/29 (-1.76)
ig_structural	49/61	46/61 (-2.79)
ig_structural_hires	20/20	19/20 (-3.03)
Gillis Collection	38/45	43/45 (-5.58)

Table 3**Examples of the most-favored contact patterns**

The indices of contact block pairs (illustrated in Fig.4) are shown in the column “Contact Patterns.”

Interaction	Contact Pattern	Notes
π - π	17-17i	offset orthogonal
	17-17ii	
	18-18i	offset orthogonal
	18-18ii	offset orthogonal
	17-18i	offset parallel
	17-18ii	offset parallel
cation- π	16-18i	orthogonal
	16-18ii	orthogonal
	16-16	orthogonal
	13-18	parallel
	13-16	parallel
cyclic side-chain	18-19	offset orthogonal
	16-19	parallel
disulfide bond	8-8i	
	8-8ii	
hydrophobic	15-15	contact at rear, two O apart
	12-19	
	12-12i	offset parallel
	12-12ii	contact at rear
	10-18	(10) rear in contact w/ (18) side
	12-18	offset parallel
	12-15	contact at rear
	12-16	parallel
	11-13i	(13) side in contact w/ (11) head
	11-13ii	(13) head points to(11) head
electrostatic	14-18	N-N
	15-19	parallel
	14-15	O-O
	2-17	hydrogen bond
	10-11	N points to midpoint of two O