RICE UNIVERSITY
A series of advanced scoring functions in ranking protein structures

By
Tianqi Ma

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy
APPROVED, THESIS COMMITTEE

Jianpeng Ma
Professor of Bioengineering, Rice University
and Lodwick T. Bolin Professor of
Biochemistry, Baylor College of Medicine

Peter Nordlander
Wiess Professor of Physics and Astronomy,
Rice University

Ching-Hwa Kiang
Associate Professor of Physics and
Astronomy and Bioengineering, Rice
University

HOUSTON, TEXAS
May 2021
ABSTRACT

A series of advanced scoring functions in ranking protein structures
by
Tianqi Ma

Designing an efficient scoring function is one of the most challenging tasks in computational biology. A good potential functions or scoring functions can help rank protein structures models, guide the search and identify possible solutions. This is important for protein structure prediction. A lot of work has been done in this area but none of them has achieved the desired result. It is therefore urgently needed to develop good scoring functions to accelerate the process.

In this thesis, I will present several novel empirical potential functions and scoring functions to address this problem. First, an upgraded version of previous work OPUS-PSP, named OPUS-DOSP. A distance related term is added to the potential function and the performance is improved. Second, a non-traditional scoring function, named OPUS-CSF is developed. This scoring function did not use the traditional Boltzmann formula but constructed a native configuration distribution table instead. This scoring function outperformed the previous work, OPUS-DOSP. Thirdly, two scoring functions combining the features of previous two scoring functions are developed. OPUS-SSF and OPUS-Beta are their names. These two scoring functions yield the best result so far and are promising in this area.
The effectiveness of these scoring functions is tested in various decoy sets generated from native structures. In the traditional benchmarks like ROSETTA, ig_structure, fisa_casp3, MOULDER and so on, OPUS-DOSP, OPUS-CSF, OPUS-SSF are performing better than previous works. In beta-prediction benchmarks Beta916 and Beta1452, OPUS-Beta also outperformed existing methods. Therefore, these scoring functions seem to be promising and useful in this field.
Acknowledgments

I would like to thank my advisor, Dr. Jianpeng Ma and our long-term collaborator Dr. Qinghua Wang for their patience, enthusiasm in science and continuous support of my study and research. As one of the best and smartest scientists, Dr. Ma always points out the keys to the problem which guide the research direction to a correct way. Without his enlightenment, the new ideas in this thesis would never come into reality.

My sincere thanks also go to the rest of my thesis committee: Dr. Kiang and Dr. Nordlander for their encouragement and insightful advices.

I would also like to thank all my current and former colleagues in Rice University and Baylor College of Medicine, Dr. Fengyun Ni, Dr. Junqing Du, Dr. Qingqing Xie, Dr. Tianwu Zang, Dr. Zhenwei Luo, and other colleagues of Dr. Ma and Dr. Wang’s lab for the stimulating discussions and enjoyable time together. My thanks also go to Dr. Gang Xu in Fudan University, without his inspiring ideas and programming skills this work cannot be done.

I want to thank Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computing resources and technical supports.

And lastly but most importantly, I would like to thank my parents, Shicheng Ma and Lianjiao Wan, my girlfriend Zhi Yan, my cat Cheese and all my friends at Rice University and Baylor College of Medicine. Without their love and support, I cannot
go through the hard times in Covid-19 lockdown and this thesis would have never been possible.
Contents

Acknowledgments .......................................................................................................................... iv

Contents ........................................................................................................................................ vi

List of Figures ................................................................................................................................. viii

List of Tables ................................................................................................................................. xi

List of Equations ............................................................................................................................. xiv

Nomenclature ................................................................................................................................. xv

Chapter 1 ......................................................................................................................................... 16

Introduction to potential functions and scoring functions .......................................................... 16

1.1. Protein Structure determination ............................................................................................. 16

1.1.1. Protein Structures and folding process .............................................................................. 17

1.2. Computational tools in predicting protein structures ........................................................... 18

1.2.1. Introduction ......................................................................................................................... 18

1.2.2. ab initio modeling ................................................................................................................ 19

1.2.3. Template-based modeling .................................................................................................. 21

1.3. Physics-based potentials and Statistical potentials .............................................................. 21

1.3.1. Physics-based potentials .................................................................................................... 22

1.3.2. Statistical potentials .......................................................................................................... 22

Chapter 2 ......................................................................................................................................... 24

OPUS-DOSP: A Distance- and Orientation-Dependent All-Atom Potential Derived from Side-Chain Packing ...................................................................................................................................................................................... 24

2.1. Orientation term of OPUS-DOSP ............................................................................................ 25

2.2. Definition of relative orientation bins ..................................................................................... 28

2.3. Definition of distance term in OPUS-DOSP .......................................................................... 31

2.4. Energy terms of OPUS-DOSP ............................................................................................... 31

2.4.1. Orientation energy term ...................................................................................................... 31

2.4.2. Distance energy term .......................................................................................................... 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.3. Construction of OPUS-DOSP energy and auxiliary function</td>
<td>33</td>
</tr>
<tr>
<td>2.5. Results of OPUS-DOSP</td>
<td>37</td>
</tr>
<tr>
<td>2.5.1. Overall performance of OPUS-DOSP</td>
<td>37</td>
</tr>
<tr>
<td>2.5.2. Results on testing the contact-range-dependent scheme and cutoff values</td>
<td>39</td>
</tr>
<tr>
<td>2.6. Discussion</td>
<td>40</td>
</tr>
<tr>
<td>2.6.1. Justification of auxiliary function</td>
<td>40</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>45</td>
</tr>
<tr>
<td>OPUS-CSF: A C-atom-based scoring function for ranking protein structural models</td>
<td>45</td>
</tr>
<tr>
<td>3.1. Introduction to OPUS-CSF</td>
<td>45</td>
</tr>
<tr>
<td>3.2. Constructing OPUS-CSF</td>
<td>46</td>
</tr>
<tr>
<td>3.2.1. Construction of CND lookup table</td>
<td>46</td>
</tr>
<tr>
<td>3.2.1. Construction of OPUS-CSF score</td>
<td>48</td>
</tr>
<tr>
<td>3.3. Results of OPUS-CSF</td>
<td>49</td>
</tr>
<tr>
<td>3.4. Discussion of OPUS-CSF</td>
<td>51</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>58</td>
</tr>
<tr>
<td>Combination of previous scoring potentials: OPUS-SSF and OPUS-Beta</td>
<td>58</td>
</tr>
<tr>
<td>4.1. Introduction to OPUS-SSF</td>
<td>58</td>
</tr>
<tr>
<td>4.2. Results and Discussion of OPUS-SSF</td>
<td>59</td>
</tr>
<tr>
<td>4.3. Introduction to OPUS-Beta</td>
<td>62</td>
</tr>
<tr>
<td>4.4. Construction of OPUS-Beta</td>
<td>64</td>
</tr>
<tr>
<td>4.4.1. Construction of lookup table</td>
<td>64</td>
</tr>
<tr>
<td>4.4.2. Construction of OPUS-Beta score</td>
<td>65</td>
</tr>
<tr>
<td>4.4.3. Results of OPUS-Beta</td>
<td>72</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>75</td>
</tr>
<tr>
<td>Conclusion</td>
<td>75</td>
</tr>
<tr>
<td>Reference</td>
<td>77</td>
</tr>
<tr>
<td>Appendix A</td>
<td>84</td>
</tr>
<tr>
<td>Appendix B</td>
<td>90</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2-1 - Definition of 19 rigid body blocks. (a): Illustration of block types. We put 19 block types into 9 block classes. Block classes I-III and block class VI are line blocks, block classes IV, V, VII, VIII and IX are plane blocks. (b): Decomposition using 19 blocks as basis, the blocks are circled in dashed lines.........................................................................................................................27

Figure 2-2 - The definition of relative orientation. Assuming block types $a$ and $b$ are in contact, then $r_{ab}$ and $r_{ba}$ are the relative direction vectors of block types $a$ and $b$, $\psi_{ab}$ is the inter-rotation angle along the axis connecting the origin point $r_{ab}$ and $r_{ba}$ of the two blocks.................................................................29

Figure 2-3 - Definition of relative direction bins. (a): Relative direction bins (b): The direction bins viewed on a Mercator projection.................................30

Figure 2-4 Schematic diagrams of $-\log(x)$ and $f(x)$..................................................................................................................................................36

Figure 2-5 Illustration of posterior distribution .................................................................42

Figure 2-6 The performance of OPUS-DOSP with different cut-off values on 11 decoy sets. (a) is the successfully recognized total number of native structures in all 11 decoy sets and the mean Z-score of all 11 decoy sets. (b) is the details of cases in (a)........................................................................................................................................44

Figure 3-1 Local molecular coordinate system in OPUS-CSF..............................48

Figure 3-2 The histogram of standard deviations of the coordinate components in the CND lookup table for 5-residue segment case. The distribution peaks at a very small value of standard deviation indicating that the coordinate components of the 1st and 5th mainchain C (carbonyl) are clustered in a narrow distribution, that is, the configurational distributions of the 5-residue peptide segments are narrow. In addition, the average value of the standard deviation is 1.20 Å .........................................................................................................................52

Figure 3-3 The population distribution of CSF scores for 278 native structures in 11 decoy sets. The X-axis is the CSF score (per independent coordinate component variable). The Y-axis is the histogram of the population. ................53

Figure 3-4 The distribution of frequency of sequence repeating in the CND lookup table. The X-axis is the repeating frequency, and the Y-axis is the
number of sequences with particular repeating frequency. Sequences that repeat less than five times were omitted in our study. Analysis of this distribution indicates that half of the sequences repeat >26 times. The largest value of X-axis is 29,618 with one sequence, but not shown for the purpose of clarity.

Figure 4-1 Schematic illustration of the five different self-packing patterns for the self-packing energy term: (a) one parallel partner; (b) one antiparallel partner; (c) two parallel partners; (d) two antiparallel partners; (e) one parallel partner and one antiparallel partner.

Figure 4-2 Schematic illustration of different types of pairwise packing patterns between two β-residues. (a) All possible pairwise residue-residue contacts in two antiparallel strands, including four types of inter-strand pairs and two types of intra-strand pairs. AA', a hydrogen-bond-involving inter-strand pair (illustrated by L_(i-1,j+1) based on the diagram); AB', a hydrogen-bond involving residue interacting with the next hydrogen-bond-involving residue on the opposite strand (illustrated by L_(i-1,j-1) based on the diagram); aa', a non-hydrogen-bond involving inter-strand pair (illustrated by L_(i,j) based on the diagram); ab', a non-hydrogen-bond involving residue interacting with the next non-hydrogen-bond involving residue on the opposite strand (illustrated by L_(i,j-2) based on the diagram); ab, a non-hydrogen-bond involving intra-strand pair (illustrated by L_(i,j-1) based on the diagram); AB, a hydrogen-bond involving intra-strand pair (illustrated by L_(i-1,i+1) based on the diagram). (b) All possible pairwise residue-residue contacts in two parallel strands, including three types of inter-strand pairs and two types of intra-strand pairs. Aa, a hydrogen-bond involving residue interacting with a non-hydrogen-bond involving residue (illustrated by L_(i,j-1) based on the diagram); Ab, a hydrogen-bond involving residue interacting with the next non-hydrogen-bond involving residue on the opposite strand toward the C terminus (illustrated by L_(i-1,j+1) based on the diagram); Ba, a hydrogen-bond involving residue interacting with the next non-hydrogen-bond involving residue on the opposite strand toward the N terminus (illustrated by L_(i+1,j-1) based on the diagram); ab, a non-hydrogen-bond involving intra-strand pair (illustrated by L_(j-1,j+1) based on the diagram); AB, a hydrogen-bond involving intra-strand pair (illustrated by L_(i-1,i+1) based on the diagram).

Figure 4-3 The performance of different kinds of potentials on decoy set Beta916 and Beta1452. The x axis is different scoring functions. The y axis is
the number of successfully recognized native structures from their decoys. The result of RDb2C2 and bbcontact are the result of pseudo-energy derive from the produced contact map.
List of Tables

Table 2-1 The performance of different potentials on 11 decoy sets. The number in the cell is the target structures successfully recognized by OPUS-DOSP. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score is, the better. The bolded entries are the best result among all these potentials. The results indicate that OPUS-DOSP, in general, outperforms OPUS-PSP and GOAP potentials in decoy set recognition in terms of both the overall number of native structures recognized and Z-scores. Meanwhile, OPUS-DOSP with auxiliary function (column: DOSP) outperforms the case with Boltzmann term alone (column: DOSP (Boltzmann)), i.e., 249 (-5.01) vs 239 (-4.67)..........................38

Table 2-2 The performance of various forms of OPUS-DOSP on 11 decoy sets. The number in the cell is the target structures successfully recognized by OPUS-DOSP. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score is, the better. The bolded entries are the best result among all these potentials. In both cases (Boltzmann function alone and Boltzmann plus auxiliary function), the consideration of orientation and distance contributions based on contact distance range (the third column in each case) results in a better performance than the cases in which orientation or distance contributions are considered alone. Moreover, the overall performance of cases with auxiliary functions (all three columns of DOSP) outperform their corresponding cases without the auxiliary function (all three columns of DOSP (Boltzmann))..........................43

Table 3-1 The results of other potentials come from the GOAP paper. The numbers of targets, with their native structures successfully recognized by various potentials, are listed in the table. The numbers in parentheses are the average Z-scores of the native structures. The larger the absolute value of Z-score, the better. Out of the total 278 targets in 11 decoy sets, OPUSCSF5(5 residue segment) recognized 244 and OPUS-CSF (combined segment length) recognizes 257 native structures from their decoys. The bold number in each row indicates the best one among all the potential functions for that particular decoy set (if the numbers of targets are the same, the bold face entries are those having the better Z-scores)..........................50

Table 3-2 The correlation coefficient of a decoy set is the average coefficient of all targets in that decoy set. In calculating the correlation coefficients, the
native structure was excluded. OPUS-CSF has comparable average correlation coefficient with other two potentials. The bold number in each row indicates the best one among the three potential functions for that particular decoy set. For OPUS-CSF, only those results for the combined segment case are listed.

Table 3-3 Success numbers are the numbers of native structures that OPUS-CSF correctly recognized from the decoys. Numbers in parentheses (278) are the total number of native structures (or targets) in 11 decoy sets. The Z scores are the calculated for the CSF scores of the native structures with respect to their decoys. Coverage means the ratio between the number of segments available in CND lookup table and the number of total segments of a target sequence. The table shows the average coverage among 278 targets in 11 decoy sets. Unknowns are the numbers of target sequences that have <20% of coverage. For these sequences, OPUS-CSF is not applicable. Note, 5-residue case does not have sequence classified as unknown, while 7-residue case, for example, has 41 out of 278 sequences not applicable for OPUS-CSF. The number of unknown increases slightly as the length of segment increases. Note, in the combined segment case, the longer segments may make no contribution to the CSF score if they are unknowns. Since the 5-residue segment case has no unknowns, it guarantees OPUS-CSF applicable to all target sequences even in rare ones that all longer segments are regarded as unknown.

Table 3-4 Num_above5 is the number of sequence segments which occur at least five times in PDB. Num_all shows the total number of sequence segments in PDB. The ratio decreases as the length of segments increases.

Table 4-1 Definition of anchoring points.

Table 4-2 The numbers of protein targets in the decoy set, with their native structures successfully recognized by OPUS-CSF and OPUS-SSF are listed in the table. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score, the better. Out of totally 603 protein targets in 15 decoy sets, OPUS-SSF recognized 571 native structures with an average Z-score of −5.46, both values are better than that of OPUS-CSF.

Table 4-3 Average Pearson’s Correlation Coefficients of.

Table 4-4 Average RMSD Values of OPUS-CSF and OPUS-SSF on 15 Decoys Sets. The numbers for each decoy set are the average RMSD values of recognized
structures in that decoy set by two scoring functions. The numbers in the last row are the weighted average numbers of all decoy sets. The value “0 indicates that all native structures were successfully found in that decoy set. OPUS-SSF outperformed OPUS-CSF in every decoy set.
List of Equations

Equation 2-1 Orientation energy function built by Boltzmann formula ..........31
Equation 2-2 Distance energy function built by Boltzmann relationship .......32
Equation 2-3 Reference contact pair numbers..................................................33
Equation 2-4 Energy function built by Boltzmann formula ..........................35
Equation 2-5 The auxiliary function..................................................................36
Equation 2-6 OPUS-DOSP potential ..................................................................37
Equation 2-7 The posterior distribution ..............................................................41
Equation 4-1 Construction of OPUS-Beta score ..............................................65
Equation 4-2 Main chain term ..........................................................................66
Equation 4-3 Self-packing term .......................................................................67
Equation 4-4 Pairwise Packing term .................................................................69
Equation 4-5 Lattice term .................................................................................71
Equation 4-6 Side Chain Block Contact term ..................................................72
# Nomenclature

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP</td>
<td>Potential based on sidechain packing</td>
</tr>
<tr>
<td>DOSP</td>
<td>Distance- and orientation-dependent potential</td>
</tr>
<tr>
<td>DOSP</td>
<td>derived from sidechain packing</td>
</tr>
<tr>
<td>CSF</td>
<td>C-atom-based scoring function</td>
</tr>
<tr>
<td>SSF</td>
<td>side-chain-inclusive scoring function</td>
</tr>
<tr>
<td>GOAP</td>
<td>Generalized orientation- and distance-dependent all-atom potential</td>
</tr>
</tbody>
</table>
Introduction to potential functions and scoring functions

1.1. Protein Structure determination

Proteins are one of the most important complex molecules in the biological processes. For example, muscles, tissues, and organs of human body are all based on different kind of proteins. A thorough study of proteins is therefore in urgent need. A protein, usually a chain-like molecules, will be in the form of a random, extended shape after being assembled. However, it can fold itself into a ‘native structure’ that is solely dependent on the amino acid sequence. This folding process, directed by physics principles and thermodynamics, has attracted thousands of most talented scientists in this area.

The structures of proteins are highly related to their functions, therefore studying protein structures is particularly important in modern biology. A
misfolded protein can lead to severe diseases\textsuperscript{1} like Creutzfeldt–Jakob disease, Alzheimer’s disease, Parkinson’s disease and lots of body disorders\textsuperscript{2-5}. Determining protein structures is therefore a major research field in modern computational biology. In the following sections, we are going to review the basic knowledge of protein structures, potential functions and scoring functions.

1.1.1. Protein Structures and folding process

Scientists divide protein structure into four levels: primary, secondary, territory and quaternary\textsuperscript{6}. In nature, there are 20 types of amino acids to form a protein, and the primary structure of a protein is the amino acid sequence forming the protein structure. Protein secondary structure provides more information of protein 3-D structure. Two major secondary structures: alpha-helix and beta-sheet can shed light on local conformations of proteins. For territory and quaternary structures, territory structure usually reflects the information of a single structure, while quaternary structures focus on how proteins domains interact with each other when they are assembled into a large biological complex.

The protein folding process is therefore a big mystery for scientists in decades. Following thermodynamics rules, scientists believe that structures with lowest free energy correspond to native structures\textsuperscript{7}. However, scientists estimated that a regular-sized protein will have more than $10^{143}$ conformations which means visiting every state to find a native structure will take an astronomical amount of time. This is not the truth as in nature a protein will be folded in microseconds, even in milliseconds\textsuperscript{8}. This fact is often called ‘Levinthal’s paradox’\textsuperscript{8}.
In regards to this paradox, Levinthal himself made a famous statement:\footnote{9}:

“...if protein folding is sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide; this suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process.”

Later experiments supported his statement. More specifically, a ‘funnel-like energy landscape’\footnote{10-13} will be the guiding direction of folding, where the bottom of the funnel represents the native state, the local minima represent the intermediate states.

Because the folding process is a such complicated problem, just determine the protein structure by experimental methods is not enough, computational methods has become more and more important. This field has been a major research interest and I will simply review the basic of this problem in the following section.

1.2. Computational tools in predicting protein structures

1.2.1. Introduction

Predicting a three-dimensional protein structure by computational methods has a lot of advantages. The rapidly developing supercomputers made some impossible solutions become true. Calculating thousands of proteins at the same time is not an illusion. Computational methods are usually more economic and
faster to be standardized and once a new effective method is fully developed, we can simultaneously predict thousands of structures on the same time on supercomputers, which make the process of discovering more protein structures become faster and easier.

Although finding the native structure solely from the amino acid sequences right now still remain unsolved, we can divide the current methods to get rough solutions into two ways: \textit{ab initio} modeling\textsuperscript{14-17} and template-based modeling\textsuperscript{18-19}. In the first approach, we start solely from the amino acid sequence and the target can be represented in three ways: atomic\textsuperscript{20}, coarse-grained\textsuperscript{21-23}, fragment-based models\textsuperscript{24}. In the latter approach, sequence alignment, or homology modeling\textsuperscript{25-26} and fold recognition(threading)\textsuperscript{27-28} techniques are applied in the process of forming the target structure from a known template. Then multiple optimization processes\textsuperscript{14-15,17,29-31} are conducted based sophisticated scoring functions that can rank protein structures. The following sections will introduce these techniques.

\textbf{1.2.2. \textit{ab initio} modeling}

\textit{ab initio} is a Latin word, which means ‘from beginning’. In this approach, a basic assumption is that proteins will spontaneously fold into the native state with global minimum free energy\textsuperscript{10,32-34}. Then the folding problem can be treated as an optimization problem, the target structure can be obtained via computer simulations. Basically, the \textit{ab initio} modeling can be classified into three types by the resolution of models used in simulation:
All atom: in this resolution, a protein structure model is defined as a collection of atoms connected by chemical bonds. Atom-atom interaction are represented functions with parameters. Quantum mechanics and quantum chemistry can help determine these parameters. Usually, these functions include bond (2-atoms), angle (3-atoms), dihedral (4-atoms), Van der Walls (2-atoms, short-range) and Coulomb (2-atoms, long-range), etc. Also, according to the way of dealing with solvents, all-atom simulations can be classified into explicit solvent and implicit solvent simulations.

Coarse-grained: in this resolution, we use fewer atoms to represent a protein structure model, usually the main-chain or side-chain atoms. Sometimes even simplify the model into residue beads or a mixture of beads and real atoms. The advantages and disadvantages are quite obvious: we can reduce the simulation dimension and simulation time, but we may lose important information of atom-atom interactions.

Fragment-based: in this resolution, proteins are regarded as fragments connected by flexible chains. These fragments will be treated as rigid bodies and they come from libraries that contain stable and frequently used fragments in known proteins. During this process, numerous of combination of fragments are tested and ranked by scoring functions. High ranking conformations will be further optimized and refined by slightly change the shape of fragments.
1.2.3. Template-based modeling

The name speaks for this method itself; template-based modeling will use a ‘template’ as the starting point of modeling. The templates come from existing protein structures. Two major approaches are used in template-based modeling: homology modeling and threading. In the first approach, we align and compare the sequences in the database. Structures with similar sequences will be the initial point of modeling. A basic assumption of homology modeling is similar sequences will fold to similar structure and this assumption has shown to be true in many systems. Threading is a technique that can deal with non-homology modeling problems. In homology modeling we are aligning sequences, while in threading, territory structures are aligned. In this process, statistical scoring functions are used for calculating matching possibility.

1.3. Physics-based potentials and Statistical potentials

As we mentioned before, no matter ab initio modeling or template-based modeling requires a scoring function that can rank protein structures by the similarity to native structure. It is therefore urgent to have high-quality scoring functions, and this is the main topic of my thesis. In the following sections, I will briefly introduce the two major types of scoring functions, i.e., physics-based potentials and statistical potentials.
1.3.1. Physics-based potentials

Physics-based potentials use quantum mechanics and other physics-based methods to calculate parameters and rank protein structures. They are widely used in \textit{ab initio} modeling\textsuperscript{20,35-37,46} and can also be classified into all-atom potentials, such as force-fields\textsuperscript{47-49} and coarse-grained potentials like MARTINI\textsuperscript{42,50}, UNRES\textsuperscript{51-52} and OPEP\textsuperscript{53}. There are two biggest advantages of physical-based potentials: first, they can guide the whole folding process, not only ranking the result. Second, they can calculate the energy landscape discussed in previous section to understand the conformation change and protein-protein interactions. However, physics-based potentials usually require much more computing time, and the result is generally worse than statistical potentials, which limits the usage of them in real life application\textsuperscript{54-63}.

1.3.2. Statistical potentials

Statistical potentials, or empirical potentials, also known as knowledge-based potentials, make use of \textit{a priori} knowledge from protein data bank to score the unknown structures. The key of building a successful statistical potential is extract essential features from protein structures, like information of a specific atom\textsuperscript{64}, pair-wise distance between atoms/residues\textsuperscript{65-67}, dihedral angles\textsuperscript{68-69}, orientation dependence of rigid body blocks\textsuperscript{70} and distance/orientation dependence of rigid body triplets\textsuperscript{71}, etc. Many statistical learning methods are used in training the parameters, including but not limited to, Bayesian network\textsuperscript{69,72}, neural network\textsuperscript{73} and so on. They can be classified into all-atom\textsuperscript{65-66,70-71,74-79} or coarse-
grained potentials\textsuperscript{59, 64, 80-89}, just like the classification in force-fields potential.

Statistical potentials are widely used in template-based modeling and non-atomic \textit{ab initio} modeling and in refinement of protein structures, which make them an especially important research area in computational biology.

In this thesis, I am focusing on developing a series of novel advanced statistical scoring functions. These functions outperformed previous work and combines the advantages of all-atom and coarse-grained potentials. In the rest of this thesis, I will introduce a distance-dependent and orientation-dependent statistical potential, OPUS-DOSP in Chapter 2, a non-traditional coarse-grained scoring function, OPUS-CSF in Chapter 3, two combination of previous work, OPUS-SSF and OPUS-Beta in Chapter 4.
In this chapter, I will discuss the OPUS-DOSP potential. OPUS-DOSP potential is a distance-dependent and orient-dependent all-atom statistical potential. We will introduce the orientation term, the distance term, and the auxiliary function used in OPUS-DOSP. The orientation term of OPUS-DOSP is based on the work of OPUS-PSP. OPUS-PSP is an orientation-dependent statistical potential based on side-chain packing (PSP). It combines the advantage of all-atom potentials and residue-level potentials. The key feature of OPUS-PSP is describing the side-chain packing interacting patterns described by a unique basis set of rigid-body blocks. OPUS-PSP constructed a basis set by decomposing the side chain
structures of 20 amino acid residues into 19 block types. Then OPUS-PSP collected the statistics of block contact patterns in a non-redundant database to capture the essential information of side-chain contact patterns and packing orientation information without losing too much conciseness, which is a disadvantage of coarse-grained approaches. In OPUS-PSP, main-chain-main-chain interactions, especially main-chain hydrogen-bonding interactions are not considered. This potential fully focuses on side-chain packing patterns. Also, it does not explicitly calculate the solvation effect.

2.1. Orientation term of OPUS-DOSP

The orientation term is same as the definition of OPUS-PSP\textsuperscript{70}. 20 amino acids were decomposed into 19 rigid body blocks. The blocks are shown in Figure 2-1a and how to decompose the sidechain of 20 amino acids is shown in Figure 2-1b.
Figure 2-1- Definition of 19 rigid body blocks. (a): Illustration of block types. We put 19 block types into 9 block classes. Block classes I-III and block class VI are line blocks, block classes IV, V, VII, VIII and IX are plane blocks. (b): Decomposition using 19 blocks as basis, the blocks are circled in dashed lines.

There are several advantages of decomposing the residues in this way: (a) the atoms in one block are all connected and belong to the same residue; (b) we assume the connected atoms form a rigid body; (c) for the convenience of analyzing
contact patterns, all heavy atoms treated in the same plane except block type 19.

2.2. Definition of relative orientation bins

To determine the contact pattern, we need to define the key variables. The contact blocks are defined as if the distance between any atoms in the two blocks are less than 5Å, them are in contact. In OPUS-DOSP, we only consider contact blocks. As we shown in Figure 2-2, the relative orientation between two contact blocks \( a \) and \( b \) is defined by \( \Omega_{ab} \), which contains three parts, two relative direction vectors \( \mathbf{r}_{ab} \) and \( \mathbf{r}_{ba} \) connect the origin points of the two blocks in the local coordinate system of \( a \) and \( b \), and an inter-rotation angle \( \psi_{ab} \) along the axis in the direction of two origins. These three variables determine the whole contact pattern.

In order to quantitively analyze the pattern, the relative direction vectors \( \mathbf{r}_{ab} \) and \( \mathbf{r}_{ba} \) are coarse-grained into 26 bins base on local molecular coordinate systems located in block \( a \) and \( b \). The inter-rotation angle \( \psi_{ab} \) is also coarse-grained into bins for simplicity. The details of building the local molecular system and determine angle \( \psi_{ab} \) will be discussed in Appendix A. The relative bins are shown in Figure 2-3a and Figure 2-3b. The direction vectors are computed and assigned to a direction bin, then the inter-rotation angle \( \psi_{ab} \) is computed, thus fully determining the contact pattern of two blocks. Figure 2-1 (b) and Figure 2-3 comes from Figure 1 and Figure 3 in OPUS-PSP, as we share the same definition.
Figure 2-2 - The definition of relative orientation. Assuming block types $a$ and $b$ are in contact, then $r_{ab}$ and $r_{ba}$ are the relative direction vectors of block types $a$ and $b$, $\psi_{ab}$ is the inter-rotation angle along the axis connecting the origin point $r_{ab}$ and $r_{ba}$ of the two blocks.
Figure 2-3 - Definition of relative direction bins. (a): Relative direction bins (b): The direction bins viewed on a Mercator projection
2.3. Definition of distance term in OPUS-DOSP

In OPUS-DOSP, compared to OPUS-PSP, we added a distance term. The distance term is defined to be the distance between two origin points of two blocks. The original OPUS-PSP do not take distance into consideration, without distance term, OPUS-PSP cannot distinguish the different cases of contact pattern with different block-block distances. That is the starting point of implementing OPUS-DOSP.

2.4. Energy terms of OPUS-DOSP

2.4.1. Orientation energy term

The orientation Boltzmann term of DOSP potential is defined by relation:

$$E_{\text{Boltz,ori}} = -k_B T \log \frac{p_{\text{obs}}(a, b, \Omega_{ab})}{p_{\text{ref}}(a, b, \Omega_{ab})}$$

Equation 2-1 Orientation energy function built by Boltzmann formula

$p_{\text{obs}}(a, b, \Omega_{ab})$ is the probability of a certain contact state $\Omega_{ab}$ for a contact block pair $a$ and $b$, defined as $p_{\text{obs}}(a, b, \Omega_{ab}) = N_{\text{obs}}(a, b, \Omega_{ab})/N_{\text{total}}^{\text{obs}}$. Here $N_{\text{obs}}(a, b, \Omega_{ab})$ is the number of observed contact state $\Omega_{ab}$ for contact block pair $a$ and $b$ and $N_{\text{total}}^{\text{obs}} = \sum_{a,b} N_{\text{obs}}(a, b, \Omega_{ab})$. Term $p_{\text{ref}}(a, b, \Omega_{ab})$ is the probability of a specific contact state $\Omega_{ab}$ in the reference state. Similarly, it is defined as $p_{\text{ref}}(a, b, \Omega_{ab}) = N_{\text{ref}}(a, b, \Omega_{ab})/N_{\text{total}}^{\text{ref}}$ with $N_{\text{ref}}(a, b, \Omega_{ab}) = \sum_{a,b} N_{\text{ref}}(a, b, \Omega_{ab})$. 


\( p(\mathbf{r}_{ab})p(\mathbf{r}_{ba})p(\psi_{ab}) N_{\text{total}}^{\text{ref}} \). Here we define \( p(\mathbf{r}_{ab}) \) is the probability of total contact pairs that have relative direction vectors \( \mathbf{r}_{ab} \) in all contact block pairs. The definitions of \( p(\mathbf{r}_{ba}) \) and \( p(\psi_{ab}) \) are similar. \( N_{\text{total}}^{\text{ref}} = \sum_{a,b} N_{\text{ref}}^{\text{ref}}(a, b, \Omega_{ab}) \). All the variables \( \mathbf{r}_{ba}, \mathbf{r}_{ab} \) and \( \psi_{ab} \) are treated as independent variables, which is a change compared to original OPUS-PSP.

### 2.4.2. Distance energy term

The distance energy term is constructed like Equation 2-2

\[
E_{\text{Boltz, dist}} = -k_B T \log \frac{p^{\text{obs}}(a, b, r)}{p^{\text{ref}}(a, b, r)}
\]

**Equation 2-2- Distance energy function built by Boltzmann relationship**

The term \( p^{\text{obs}}(a, b, r) \) is the probability of a certain contact block pair \( a \) and \( b \) with specific distance \( r \), defined as \( p^{\text{obs}}(a, b, r) \equiv N^{\text{obs}}(a, b, r) / N_{\text{total}}^{\text{obs}} \). Here \( N^{\text{obs}}(a, b, r) \) is the number of observed contact pair with specific distance \( r \) for contact block pair \( a \) and \( b \) and \( N_{\text{total}}^{\text{obs}} = \sum_{a,b} N^{\text{obs}}(a, b, r) \). The term \( p^{\text{ref}}(a, b, r) \) is the probability of a specific distance \( r \) for a contact block pair \( a \) and \( b \) in reference state defined as \( p^{\text{ref}}(a, b, r) \equiv N^{\text{ref}}(a, b, r) / N_{\text{total}}^{\text{ref}} \). \( N^{\text{ref}}(a, b, r) \) is the number of a specific distance \( r \) between contact block pair of \( a \) and \( b \) in reference state. We use the similar techniques to DFIRE\textsuperscript{78}, the distribution of reference state is shown in Equation 2-3.
\[
N_{\text{ref}}(a, b, r) = \frac{\Delta r}{r_{\text{cut}}} p(a, b) N_{\text{total}}^{\text{obs}}
\]

Equation 2-3 Reference contact pair numbers

The term \( p(a, b) \) follows quasi-chemical assumption \( p(a, b) \approx \chi_a \chi_b \) and \( \chi_a \) is the mole fraction of block type \( a \) and \( N_{\text{total}}^{\text{ref}} = \sum_{a, b} N_{\text{ref}}^a N_{\text{ref}}^b \). The cutoff value is set to be 15 Å and \( \Delta r \) is 2 Å for \( r < 2 \, \text{Å} \), 0.5 Å for \( 2 \, \text{Å} < r < 8 \, \text{Å} \), and 1 Å for \( 8 \, \text{Å} < r < 15 \, \text{Å} \). It is assumed that the reference state obeys the uniform distribution so that the quasi-chemical assumption holds.

2.4.3. Construction of OPUS-DOSP energy and auxiliary function

In OPUS-DOSP, the contributions of orientation and distance parts may change when the distance varies. When the distance between two contacting blocks is either too small (less than \( r_1 = 3.7 \, \text{Å} \)) or too large (larger than \( r_2 = 10 \, \text{Å} \)), we only consider the orientation term, otherwise in the middle-distance range we only used the distance term. In the short distance case, the dominating factor is orientation patterns for contact block patterns. In the intermediate distance case, the distance dependence will maximize its effect. In the large distance case, the sole usage of orientation term will be explained later in the subsection. The arrangement of orientation term and distance term is called a contact-range-dependent combination scheme of energy contributions. It is possible that. In this work, only one energy term in each distance range will be used to avoid optimizing that weight.
function. Another feature of OPUS-DOSP is that contact pairs connected by chemical bonds, which only happens intra-residue pairs are discarded.

2.4.3.1. Discussion of contact-range-dependent scheme

In previous section, we discussed a contact-range-dependent scheme. Since some blocks probably have more physical contacts than the others depending on the packing pattern, for simplicity, the distance between the origins of two contacting blocks is used as a parameter to distinguish the packing pattern. We assume when two contacting blocks are in close distance (distance < \( r_1 \)), the orientation contribution dominates. This is because, in close distance, the two blocks are likely to be closely contacted and the remaining freedom to move between the two blocks are the relative orientation rather than the distance. Thus, we only consider the orientation contribution in close distance range. When the distance between two blocks is in the intermediate range (\( r_1 < \text{distance} < r_2 \)), two blocks can have much more various contact patterns, the blocks are more freely to rotate. In this case, distance term seems to be more sensitive than the orientation term. We therefore only consider the distance term energy function. When the distance between two blocks is large (distance > \( r_2 \)), the two blocks need to be in specific position in order to have one or a few atoms in contact. Therefore, the orientation term is again sensitive. We thus only use orientation contribution in this case. The actual values of \( r_1 = 3.7\text{Å} \) and \( r_2 = 10\text{Å} \) were empirically determined. The results for OPUS-DOSP performance using orientation contribution alone, distance contribution alone, and
contact-range-dependent combination of the two will be discussed in the following results section.

2.4.3.2. Auxiliary function

Almost all empirical potentials are established based on Boltzmann formula,

\[ E_{\text{Boltz}} = -k_B T \log \frac{p^{\text{obs}}}{p^{\text{ref}}} \]

Equation 2-4 Energy function built by Boltzmann formula

Where \( k_B T \) is the Boltzmann constant and temperature, \( p^{\text{obs}} \) is the probability of specific contact pattern in the observed state, \( p^{\text{ref}} \) is the probability of specific contact pattern in the reference state. In this thesis, we found that Boltzmann formula may underestimate the influence of extreme cases. The reason is that we only have limited structures in the non-redundant structure database we used to construct the potential, which makes the packing patterns distribution in that database deviate from Boltzmann distribution due to limited sampling. The consequence for the limited sampling of packing pattern is that the extreme cases, the ratio of \( p^{\text{obs}} / p^{\text{ref}} \) is exceptionally large (extremely favorable case) or exceedingly small (extremely unfavorable case), can produce more reliable results, and we want to increase the weight of these cases in building up OPUS-DOSP potential. Therefore, an auxiliary function \( f(x) \) is added to the original Boltzmann relationship to increase the weight of extreme cases. This auxiliary function is illustrated in Figure 2-4 and the function form is shown in Equation 2-5. The reason
that we choose a non-continuous form function will be explained in Appendix B.

This function is clearly raising the weights of extreme cases and fit our requirement.

Figure 2-4 Schematic diagrams of $-\log(x)$ and $f(x)$

\[
f(x) = \begin{cases} 
-x & \text{if } x > 1 \\
0 & \text{if } x = 0 \\
\frac{1}{x} & \text{if } x < 1 
\end{cases}
\]

Equation 2-5 The auxiliary function
2.4.3.3. Final energy form of the OPUS-DOSP

The final form of OPUS-DOSP is a combination of Boltzmann term energy and the energy in auxiliary function form. The energy term looks like this:

\[ E_{\text{ori}}(a, b, \Omega_{ab}) = E_{\text{Boltz,ori}}(a, b, \Omega_{ab}) + f \left( \frac{p^{\text{obs}}(a, b, \Omega_{ab})}{p^{\text{ref}}(a, b, \Omega_{ab})} \right) \]

\[ E_{\text{dist}}(a, b, r) = E_{\text{Boltz, dist}}(a, b, r) + f \left( \frac{p^{\text{obs}}(a, b, r)}{p^{\text{ref}}(a, b, r)} \right) \]

**Equation 2-6  OPUS-DOSP potential**

In practice, we will setup some energy cutoffs to avoid individual extreme energy values. We defined a cutoff value \(c\), if the input value to \(f(x)\) is smaller than \(1/c\) or larger than \(c\), then we set the input to be \(1/c\) or \(c\) so that the output of auxiliary function is \(c\). In OPUS-DOSP, the cutoff value is set to be 55, and this result can be derived from the logarithm result and the detailed process can be found in Appendix B.

2.5. Results of OPUS-DOSP

2.5.1. Overall performance of OPUS-DOSP

We trained the potential based on the same 1011 proteins used to train GOAP. The performance of OPUS-DOSP was tested on 11 commonly used decoy
sets that are also used in GOAP potential. Including 4state_reduced\textsuperscript{91}, fisa\textsuperscript{44},
fisa_casp3\textsuperscript{44}, hg_structal, ig_structal, ig_structal_hires (R. Samudrala, E. Huang, and
M. Levitt, unpublished), I-TASSER\textsuperscript{65}, lattice_ssfit\textsuperscript{92-93}, lmds\textsuperscript{94}, MOULDER\textsuperscript{95} and
ROSETTA\textsuperscript{96}. The results of OPUS-PSP, OPUS-DOSP, GOAP potential are presented in
Table 1.

<table>
<thead>
<tr>
<th>Decoy sets</th>
<th>Numbers of targets</th>
<th>PSP</th>
<th>GOAP</th>
<th>DOSP (Boltzmann)</th>
<th>DOSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>7</td>
<td>7(-4.41)</td>
<td>7(-4.31)</td>
<td>5(-4.26)</td>
<td>3(-4.03)</td>
</tr>
<tr>
<td>fisa</td>
<td>4</td>
<td>3(-4.07)</td>
<td>3(-3.94)</td>
<td>4(-5.12)</td>
<td>2(-3.77)</td>
</tr>
<tr>
<td>fisa_casp3</td>
<td>5</td>
<td>5(-6.22)</td>
<td>5(-5.16)</td>
<td>4(-4.33)</td>
<td>4(-4.40)</td>
</tr>
<tr>
<td>hg_structal</td>
<td>29</td>
<td>18(-1.75)</td>
<td>22(-1.98)</td>
<td>25(-3.25)</td>
<td>27(-3.35)</td>
</tr>
<tr>
<td>ig_structal</td>
<td>61</td>
<td>22(-1.06)</td>
<td>47(-1.53)</td>
<td>61(-6.91)</td>
<td>61(-7.08)</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>20</td>
<td>15(-1.58)</td>
<td>18(-1.82)</td>
<td>20(-4.20)</td>
<td>20(-4.24)</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>56</td>
<td>45(-3.46)</td>
<td>45(-4.99)</td>
<td>56(-5.55)</td>
<td>51(-4.97)</td>
</tr>
<tr>
<td>lattice_ssfit</td>
<td>8</td>
<td>8(-6.52)</td>
<td>8(-8.53)</td>
<td>5(-4.56)</td>
<td>3(-4.46)</td>
</tr>
<tr>
<td>lmds</td>
<td>10</td>
<td>8(-5.2)</td>
<td>7(-3.54)</td>
<td>10(-5.81)</td>
<td>10(-7.43)</td>
</tr>
<tr>
<td>MOULDER</td>
<td>20</td>
<td>19(-4.62)</td>
<td>19(-3.48)</td>
<td>15(-2.99)</td>
<td>17(-4.25)</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>58</td>
<td>39(-3.17)</td>
<td>45(-3.39)</td>
<td>34(-2.93)</td>
<td>51(-4.16)</td>
</tr>
<tr>
<td>Total</td>
<td>278</td>
<td>189(-2.87)</td>
<td>226(-3.27)</td>
<td>239(-4.67)</td>
<td>249(-5.01)</td>
</tr>
</tbody>
</table>

Table 1 The performance of different potentials on 11 decoy sets. The number in the cell is the target structures successfully recognized by OPUS-DOSP. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score is, the better. The bolded entries are the best result among all these potentials. The results indicate that OPUS-DOSP, in general, outperforms OPUS-PSP and GOAP potentials in decoy set recognition in terms of both the overall number of native structures recognized and Z-scores. Meanwhile, OPUS-DOSP with auxiliary function (column: DOSP)
outperforms the case with Boltzmann term alone (column: DOSP (Boltzmann)), i.e., 249 (-5.01) vs 239 (-4.67).

From the result, it is shown that OPUS-DOSP successfully recognized 249 out 278 native structures from their decoys with the lowest average Z-score. It is worth noting that even without the auxiliary function, OPUS-DOSP recognized 239 out of 278 native structures from their decoys. Both forms of OPUS-DOSP outperformed existing methods, indicates that combining the rigid body representation of sidechains of amino acids and distance between the origin of blocks can extract essential features of contacting patterns. More specifically, in three homology modeling sets (hg_structal, ig_structal, ig_structal_hires), OPUS-DOSP performs pretty well. When the auxiliary function was added, the overall performance of DOSP was improved but the result for 4state_reduced, fisa, ITASSER and lattice-ssfit, the performance slightly dropped.

2.5.2. Results on testing the contact-range-dependent scheme and cutoff values

In the previous discussion, we introduced a contact-range-dependent scheme. In close distance and large distance range, we use orientation term only. While in middle-range, only the distance term will be taken into consideration. In this section, we will show the effectiveness of this scheme.

First, we will test the choice of our combination of orientation term and distance term: we assumed that when two contacting blocks are close or far away,
we use the orientation term only; when the two blocks are in middle-range, we only consider the distance term. Different combination of orientation term and distance terms were tested on the 11 decoy sets discussed before and the result with or without the auxiliary function is also shown in Table 2. It can be seen that the contact-range-dependent combination has better performance than orientation or distance term alone, and auxiliary function constantly improved the ability of recognizing native structures. Therefore, the final OPUS-DOSP will be the potential with contact-range-combination scheme and an auxiliary function. The actual values deciding the close distance or far-away distance is determined empirically. We have $r_1 = 3.7\text{Å}$ and $r_2 = 10\text{Å}$.

Then, the effectiveness of these cutoff values was tested. Different setup of cutoff distance values was tested. The options are zero, original values or infinite. The result is shown in Figure 2-6. It is clear that our cutoff values increased the performance of OPUS-DOSP both in decoy recognition and Z-scores.

2.6. Discussion

2.6.1. Justification of auxiliary function

The auxiliary function is not introduced in previous works, and we want to justify the usage of this function here by a simple binomial model. A block pair with contact type a and b will have 21 translation bins and $26 \times 4 \times 26 = 2704$ orientation bins. We denote $\theta$ as the probability for the a-b contact pair in a specific bin, then in practice, the precise probability $\theta$ cannot be calculated due to the
limited sampling size, we need the information of all contact pairs in native structures, which is not possible. Therefore, we can only estimate this value via sampling our training set. Suggest we have N pairs of block type a and b, the number of observations in that bin obeys Binomial distribution \((N, \theta)\), we estimate the parameter by relation \(\theta \approx \frac{y}{n}\) where \(y\) is the number observed in that bin. Then, following the Bayesian inference calculation, using a uniform prior, the posterior distribution will be a Beta distribution,

\[
p(\theta|y) \sim \text{beta}(1 + y, 1 + n - y)
\]

**Equation 2-7 The posterior distribution**

Take \((y = 9, n = 10)\), \((y = 1, n = 10)\) and \((y = 5, n = 10)\) as examples. From these data, we will estimate \(\theta = 0.9\), \(\theta = 0.1\) and \(\theta = 0.5\), respectively. The posterior distributions of parameter \(\theta\) are plotted in Figure 2-5. It is obvious that \(p(\theta|y)\) for extreme cases \((y = 9, x=0.9\) or \(y = 1, x=0.1\)) has greater probability density at \(x=0.9\) or \(x=0.1\) than non-extreme case \((y=5, x=0.5)\). Therefore, the extreme cases are more reliable than non-extreme cases. It is possible that this result heavily depends on the uniform prior, but we do not have any *in priori* information, a uniform prior is a reasonable choice. Therefore, the idea of increasing the weight of extreme cases is reasonable.

What’s more, the fact that we used a non-Boltzmann formula improved the performance inspired us to think about non-Boltzmann scoring functions, the result will be discussed in next chapter.
Figure 2-5 Illustration of posterior distribution
<table>
<thead>
<tr>
<th>Decoy sets</th>
<th>Number of targets</th>
<th>Orientation alone</th>
<th>Distance alone</th>
<th>Orientation &amp; distance</th>
<th>Orientation alone</th>
<th>Distance alone</th>
<th>Orientation &amp; distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>7</td>
<td>4(-2.97)</td>
<td>3(-3.46)</td>
<td>5(-4.26)</td>
<td>6(-4.32)</td>
<td>2(-2.26)</td>
<td>3(-4.03)</td>
</tr>
<tr>
<td>fisa</td>
<td>4</td>
<td>2(-1.37)</td>
<td>3(0.007)</td>
<td>4(-5.12)</td>
<td>1(-1.29)</td>
<td>3(-2.96)</td>
<td>2(-3.77)</td>
</tr>
<tr>
<td>fisa_casp3</td>
<td>5</td>
<td>1(-1.79)</td>
<td>2(3.42)</td>
<td>4(-4.33)</td>
<td>2(-2.19)</td>
<td>1(-1.57)</td>
<td>4(-4.40)</td>
</tr>
<tr>
<td>hg_structal</td>
<td>29</td>
<td>17(-1.18)</td>
<td>20(-2.34)</td>
<td>25(-3.25)</td>
<td>10(-0.95)</td>
<td>23(-2.85)</td>
<td>27(-3.35)</td>
</tr>
<tr>
<td>ig_structal</td>
<td>61</td>
<td>52(-3.57)</td>
<td>61(-6.54)</td>
<td>61(-6.91)</td>
<td>50(-3.08)</td>
<td>61(-7.02)</td>
<td>61(-7.08)</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>20</td>
<td>19(-3.38)</td>
<td>20(-4.13)</td>
<td>20(-4.20)</td>
<td>19(-3.10)</td>
<td>20(-4.23)</td>
<td>20(-4.24)</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>56</td>
<td>50(-5.55)</td>
<td>44(-4.06)</td>
<td>56(-5.55)</td>
<td>44(-4.41)</td>
<td>51(-4.43)</td>
<td>51(-4.97)</td>
</tr>
<tr>
<td>lattice_ssfit</td>
<td>8</td>
<td>6(-4.27)</td>
<td>6(-1.79)</td>
<td>5(-4.56)</td>
<td>4(-2.90)</td>
<td>7(-3.82)</td>
<td>3(-4.46)</td>
</tr>
<tr>
<td>lmds</td>
<td>10</td>
<td>6(-2.74)</td>
<td>8(-2.03)</td>
<td>10(-5.81)</td>
<td>7(-3.76)</td>
<td>8(-4.03)</td>
<td>10(-7.43)</td>
</tr>
<tr>
<td>MOULDER</td>
<td>20</td>
<td>13(-2.53)</td>
<td>11(-0.58)</td>
<td>15(-2.99)</td>
<td>17(-3.12)</td>
<td>14(-2.94)</td>
<td>17(-4.25)</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>58</td>
<td>16(-1.22)</td>
<td>5(4.48)</td>
<td>34(-2.93)</td>
<td>34(-2.50)</td>
<td>4(2.80)</td>
<td>51(-4.16)</td>
</tr>
<tr>
<td>Total</td>
<td>278</td>
<td>186(-3.11)</td>
<td>183(-1.98)</td>
<td>239(-4.67)</td>
<td>194(-3.01)</td>
<td>194(-3.04)</td>
<td>249(-5.01)</td>
</tr>
</tbody>
</table>

Table 2 The performance of various forms of OPUS-DOSP on 11 decoy sets. The number in the cell is the target structures successfully recognized by OPUS-DOSP. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score is, the better. The bolded entries are the best result among all these potentials. In both cases (Boltzmann function alone and Boltzmann plus auxiliary function), the consideration of orientation and distance contributions based on contact distance range (the third column in each case) results in a better performance than the cases in which orientation or distance contributions are considered alone. Moreover, the overall performance of cases with auxiliary functions (all three columns of DOSP) outperform their corresponding cases without the auxiliary function (all three columns of DOSP (Boltzmann)).
Figure 2-6 The performance of OPUS-DOSP with different cut-off values on 11 decoy sets. (a) is the successfully recognized total number of native structures in all 11 decoy sets and the mean Z-score of all 11 decoy sets. (b) are the details of cases in (a).
OPUS-CSF: A C-atom-based scoring function for ranking protein structural models

As we discussed in last chapter, using a non-Boltzmann formula in constructing the potential inspired us to think outside the box. Therefore, we developed a non-Boltzmann scoring function, OPUS-CSF to rank protein structures\(^{97}\). The performance of OPUS-CSF improved a lot comparing to previous empirical potentials, indicating possible improvement for further works.

3.1. Introduction to OPUS-CSF

In this work, unlike traditional empirical potential functions using Boltzmann formula, we developed a scoring function based on the distributions of coordinate components of mainchain C (carbonyl) atoms in native structures on a few selected
residues of small peptide segments of 5, 7, 9, and 11 residues in length. We built a configurational native distribution (CND) lookup table that contains native distributions of coordinate components. The table was built by analyzing peptide segments in the entire Protein Data Bank (PDB). In other word, the whole Protein Data Bank is our training set.

To calculate the score of a test structure, we need to compare the segment information of the structure with the same segment information stored in CND lookup table. The performance of OPUS-CSF was tested in 11 commonly used decoy sets and the result is very promising. Besides, we also calculated the correlation coefficient between TM-score and CSF score, the result is comparable to GOAP and OPUS-PSP, which is quite inspiring due to OPUS-CSF’s highly coarse-grained nature. This fact, combining with the fast speed of OPUS-CSF, made it important and promising in ranking structure models from intermediate data from experimental models and ranking models from molecular dynamics.

### 3.2. Constructing OPUS-CSF

#### 3.2.1. Construction of CND lookup table

The configurational native distribution (CND) lookup table is constructed via scanning the whole Protein Data Bank (PDB). In this process, we are scanning the polypeptide chain using a window size of length 5, 7, 9, 11 residues and step size of one residue in the entire PDB. In conducting this research, we downloaded 130,054 protein structures from PDB on June 7, 2017 via [ftp://ftp.wwpdb.org/](ftp://ftp.wwpdb.org/). Sequences
that appear less than 5 times are discarded, the number 5 is determined empirically. Also, peptide segments with poorly resolved structures such as broken bonds were not included.

The detailed process will be discussed via an example of a 5-residue segment case. First, for every 5-residue segment, we first define a local molecular system using the atom coordinates of three main-chain atoms in the third residue. The origin is set at Cα atom, the X-axis is defined as the line along the direction between Cα and C(carbonyl) atom, the Y-axis is located in Cα-C-O plane, parallel to the C-O vector projection. Z-axis is defined via right-hand coordinate system’s rule. A sketch of the coordinate system is shown in Figure 3-1.

Then for a 5-residue segment with specific residue sequence, the atom positions of mainchain C(carbonyl) atoms of 1st and 5th residue in the local coordinate system in the segment are calculated. We denote these variables as \((x_1, y_1, z_1)\) and \((x_5, y_5, z_5)\). Under our assumption, these variables are treated as six independent Gaussian variables. By scanning the entire PDB, we generate six independent Gaussian distributions, called configurational native distributions (CNDs) of 5-residue segments. Then, the mean and standard deviation of the of the distributions are calculated and stored in CND lookup table. The content in the lookup tables will be similar to a dictionary, the key is the amino acid sequence and the values are the mean and standard deviation of the coordinate components of the 1st and 5th residue.
3.2.1. Construction of OPUS-CSF score

When we finished building the CND lookup table, we can calculate the CSF score for a test structure. Once a new structure comes in hand, we will scan through the structure with window length of 5 residues. When this process ends, we calculate the absolute value of Z-score of every variable of every segments. For example, if we have 3 5-residue segments, in total there would be $2 \times 3 \times 3 =$
18 terms since every segment have two recorded residues, every residue’s main chain C atom has 3 coordinate components. The final score is called OPUS-CSF score and we assume the structure with smallest OPUS-CSF score to be the native structure.

The segments of various lengths are denoted as 5(1, 3, 5), 7(2, 4, 6), 9(1, 3, 5, 7, 9) and 11(2, 4, 6, 8, 10). Here, still consider the 5-residue segment of form 5(1, 3, 5) as an example, the first number 5 is the segment length, 1,5 in the parenthesis are the residues that we record C (carbonyl) atom positional distributions in local coordinate system, 3 is the residue on which the local coordinate system is defined. For 9(1, 3, 5, 7, 9) and 11(2, 4, 6, 8, 10), four mainchain C (carbonyl) atoms are recorded to store the positional distributions, thus totally 12 independent variables are used for these segments.

3.3. Results of OPUS-CSF

OPUS-CSF was tested on 11 commonly used decoy sets used in OPUS-DOSP and GOAP. In Table 3, we showed the performance of OPUS-CSF compared with other empirical potentials and scoring functions, to be more specifically, we tested the performance of different forms of OPUS-CSF and other empirical potentials. Overall, OPUS-CSF outperformed other scoring functions and empirical potentials by 257 out of 278 native structures from decoys with the best Z-score of -4.12 on average. What’s more, we tried to calculate the Pearson’s correlation coefficients between the CSF scores and TM-scores in the decoy set. The result is shown in Table
It is worth noting that although OPUS-CSF is just a coarse-grained scoring function, the result is better than all-atom potentials, even for the Pearson’s correlation coefficients. This inspiring result suggest that OPUS-CSF is very promising in ranking protein structures and in early stages of protein modeling.

Table 3 The results of other potentials come from the GOAP paper. The numbers of targets, with their native structures successfully recognized by various potentials, are listed in the table. The numbers in parentheses are the average Z-scores of the native structures. The larger the absolute value of Z-score, the better. Out of the total 278 targets in 11 decoy sets, OPUSCSF5(5 residue segment) recognized 244 and OPUS-CSF (combined segment length) recognizes 257 native structures from their decoys. The bold number in each row indicates the best one among all the potential functions for that particular decoy set (if the numbers of targets are the same, the bold face entries are those having the better Z-scores).
Table 4 The correlation coefficient of a decoy set is the average coefficient of all targets in that decoy set. In calculating the correlation coefficients, the native structure was excluded. OPUS-CSF has comparable average correlation coefficient with other two potentials. The bold number in each row indicates the best one among the three potential functions for that particular decoy set. For OPUS-CSF, only those results for the combined segment case are listed.

<table>
<thead>
<tr>
<th>Decoy sets</th>
<th>OPUS-PSP</th>
<th>GOAP</th>
<th>OPUS-CSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>-0.589</td>
<td>-0.694</td>
<td>-0.667</td>
</tr>
<tr>
<td>fisa</td>
<td>-0.282</td>
<td>-0.347</td>
<td>-0.552</td>
</tr>
<tr>
<td>fisa_casp3</td>
<td>-0.095</td>
<td>-0.221</td>
<td>-0.333</td>
</tr>
<tr>
<td>hg_structal</td>
<td>-0.752</td>
<td>-0.825</td>
<td>-0.803</td>
</tr>
<tr>
<td>ig_structal</td>
<td>-0.779</td>
<td>-0.865</td>
<td>-0.882</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>-0.832</td>
<td>-0.885</td>
<td>-0.901</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>-0.284</td>
<td>-0.477</td>
<td>-0.452</td>
</tr>
<tr>
<td>lattice_ssfit</td>
<td>-0.051</td>
<td>-0.058</td>
<td>-0.151</td>
</tr>
<tr>
<td>lmds</td>
<td>-0.091</td>
<td>-0.146</td>
<td>-0.342</td>
</tr>
<tr>
<td>MOULDER</td>
<td>-0.802</td>
<td>-0.886</td>
<td>-0.863</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>-0.343</td>
<td>-0.476</td>
<td>-0.391</td>
</tr>
<tr>
<td>Average</td>
<td>-0.521</td>
<td>-0.632</td>
<td>-0.624</td>
</tr>
</tbody>
</table>

3.4. Discussion of OPUS-CSF

For further analysis of this method, we analyzed the 5-residue segment data as an example. We first analyzed the standard deviations of coordinate components in 1st and 5th residues in CND lookup table in Figure 3-2. It is obvious that the standard deviation is quite small, indicating these coordinate components are clustered in a narrow distribution44,98, which provides a foundation for the success of OPUS-CSF. The average value is 1.2 Å.
Figure 3-2 The histogram of standard deviations of the coordinate components in the CND lookup table for 5-residue segment case. The distribution peaks at an exceedingly small value of standard deviation indicating that the coordinate components of the 1st and 5th mainchain C (carbonyl) are clustered in a narrow distribution, that is, the configurational distributions of the 5-residue peptide segments are narrow. In addition, the average value of the standard deviation is 1.20 Å.

Another assumption to make OPUS-CSF succeed is that the smaller the CSF score, the closer the structure to the native structure. In fact, this is an approximation since native structure will not have zero CSF score. However, Figure 3-2 suggested that the distributions of coordinate components are narrow, then assuming native structures have small CSF score is reasonable. We calculated the CSF score of 278 native structures (per coordinate component) in Figure 3-3. The average value is 0.84 and the standard deviation is 0.27, indicated that the fluctuation of native CSF scores is also exceedingly small.
The sequence repeating pattern is also studied in this thesis. We still take 5-residue segments as an example. In principle, the more times a sequence appears in the lookup table, the better statistics we will have. In the 5-residue segments, half of the sequences repeat over 26 times, the largest value is 29618. In constructing the lookup table, a tradeoff between sequence diversity and sequence repeating frequency is necessary.

We also examined the segment of different lengths in Table 6. As the length of segment increases, naturally the coverage decreases, and the ratio of the number of segments that appear more than five times to the total number of segments in PDB decreases. On the other hand, if Coverage is defined as the ratio between the number of segments available in CND lookup table and the number of total...
segments of a test sequence, the average coverage of the 11 decoy sets (in total 278 targets) decreases as the length of segment increases. If a test sequence has <20% of its segments available in the CND lookup table, that is, its coverage is <20%, it is regarded as Unknown, then the number of unknowns increase as the lengths of segments increase. The version of OPUS-CSF using 5 residues has the best result in decoy recognition, but the Z-score performance is decreased. It is possible that longer segments preserve more homology information.

There is also a cutoff value of every OPUS-CSF score. Since the probability of a sample from Gaussian distribution falls outside \((\mu - 5\sigma, \mu + 5\sigma)\), or Z-score greater than 5 is in the order of \(10^{-7}\), we set a cutoff value of 15 since we have three coordinate components.
<table>
<thead>
<tr>
<th></th>
<th>5 residues</th>
<th>7 residues</th>
<th>9 residues</th>
<th>11 residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success numbers</td>
<td>244(278)</td>
<td>218(278)</td>
<td>220(278)</td>
<td>219(278)</td>
</tr>
<tr>
<td>Z-scores</td>
<td>-3.56</td>
<td>-4.55</td>
<td>-4.62</td>
<td>-4.57</td>
</tr>
<tr>
<td>Average Coverage</td>
<td>0.971</td>
<td>0.749</td>
<td>0.712</td>
<td>0.683</td>
</tr>
<tr>
<td>Unknowns</td>
<td>0</td>
<td>41</td>
<td>45</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 5 Success numbers are the numbers of native structures that OPUS-CSF correctly recognized from the decoys. Numbers in parentheses (278) are the total number of native structures (or targets) in 11 decoy sets. The Z scores are the calculated for the CSF scores of the native structures with respect to their decoys. Coverage means the ratio between the number of segments available in CND lookup table and the number of total segments of a target sequence. The table shows the average coverage among 278 targets in 11 decoy sets. Unknowns are the numbers of target sequences that have <20% of coverage. For these sequences, OPUS-CSF is not applicable. Note, 5-residue case does not have sequence classified as unknown, while 7-residue case, for example, has 41 out of 278 sequences not applicable for OPUS-CSF. The number of unknown increases slightly as the length of segment increases. Note, in the combined segment case, the longer segments may make no contribution to the CSF score if they are unknowns. Since the 5-residue segment case has no unknowns, it guarantees OPUS-CSF applicable to all target sequences even in rare ones that all longer segments are regarded as unknown.
Figure 3-4 The distribution of frequency of sequence repeating in the CND lookup table. The X-axis is the repeating frequency, and the Y-axis is the number of sequences with repeating frequency. Sequences that repeat less than five times were omitted in our study. Analysis of this distribution indicates that half of the sequences repeat >26 times. The largest value of X-axis is 29,618 with one sequence, but not shown for the purpose of clarity.
<table>
<thead>
<tr>
<th>Residues</th>
<th>Num_above 5</th>
<th>Num_all</th>
<th>Num_above5/Num_all</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-residues</td>
<td>1766273</td>
<td>2350969</td>
<td>0.751</td>
</tr>
<tr>
<td>7-residues</td>
<td>3736778</td>
<td>9544858</td>
<td>0.391</td>
</tr>
<tr>
<td>9-residues</td>
<td>3713506</td>
<td>10262243</td>
<td>0.362</td>
</tr>
<tr>
<td>11-residues</td>
<td>3743204</td>
<td>10698802</td>
<td>0.350</td>
</tr>
</tbody>
</table>

Table 6 Num_above5 is the number of sequence segments which occur at least five times in PDB. Num_all shows the total number of sequence segments in PDB. The ratio decreases as the length of segments increases.
Chapter 4

Combination of previous scoring potentials: OPUS-SSF and OPUS-Beta

In Chapter 2 and Chapter 3, we introduced an empirical potential OPUS-DOSP and a non-Boltzmann scoring function. In this Chapter, two examples combining these two methods will be discussed, OPUS-SSF and OPUS-Beta.

4.1. Introduction to OPUS-SSF

To discuss OPUS-SSF, we need to briefly recall the empirical potential OPUS-DOSP and OPUS-CSF. In OPUS-DOSP, we decomposed the side-chain of 20 amino acid into 19 rigid body blocks. In OPUS-CSF, we constructed a configurational native distribution lookup table and calculate the absolute value of Z-scores of coordinate components of main chain carbonyl atoms. In OPUS-SSF\textsuperscript{99}, we tried to combine the information of both main-chain and side-chain and non-Boltzmann approach. To
deal with side chains, we followed the approach used in OPUS-PSP and OPUS-DOSP. Instead of storing distribution information of main chain carbonyl atom, we computed the anchoring points that may be actual or computed atom positions.

The anchoring points definition is shown in Table 7. Other setup, like CND lookup tables, the local molecular systems remain the same.

<table>
<thead>
<tr>
<th>Anchoring point 1</th>
<th>Anchoring point 2</th>
<th>Anchoring point 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLY</td>
<td>C</td>
<td>–</td>
</tr>
<tr>
<td>ALA</td>
<td>CB</td>
<td>–</td>
</tr>
<tr>
<td>SER</td>
<td>(CB + OG)/2</td>
<td>–</td>
</tr>
<tr>
<td>CYSS</td>
<td>(CB + SG)/2</td>
<td>–</td>
</tr>
<tr>
<td>VAL</td>
<td>(CB + CG1 + CG2)/3</td>
<td>–</td>
</tr>
<tr>
<td>ILE</td>
<td>(CB + CG1 + CG2)/3</td>
<td>CD</td>
</tr>
<tr>
<td>LEU</td>
<td>CB</td>
<td>(CG + CD1 + CD2)/3</td>
</tr>
<tr>
<td>THR</td>
<td>(CB + OG1 + CG2)/3</td>
<td>–</td>
</tr>
<tr>
<td>ARG</td>
<td>CB</td>
<td>(CG + CD)/2</td>
</tr>
<tr>
<td>LYS</td>
<td>CB</td>
<td>(CG + CD)/2</td>
</tr>
<tr>
<td>ASP</td>
<td>CB</td>
<td>(CG + OD1 + OD2)/3</td>
</tr>
<tr>
<td>GLU</td>
<td>(CB + CG)/2</td>
<td>(CD + OE1 + OE2)/3</td>
</tr>
<tr>
<td>ASN</td>
<td>CB</td>
<td>(CG + OD1 + ND2)/3</td>
</tr>
<tr>
<td>GLN</td>
<td>(CB + CG)/2</td>
<td>(CD + OE1 + NE2)/3</td>
</tr>
<tr>
<td>MET</td>
<td>(CB + CG)/2</td>
<td>(SD + CE)/2</td>
</tr>
<tr>
<td>HIS</td>
<td>CB</td>
<td>(CG + CE1 + NE2)/3</td>
</tr>
<tr>
<td>PRO</td>
<td>(CB + CG + CD)/3</td>
<td>–</td>
</tr>
<tr>
<td>PHE</td>
<td>CB</td>
<td>(CG + CE1 + CE2)/3</td>
</tr>
<tr>
<td>TYR</td>
<td>CB</td>
<td>(CG + CE1 + CE2)/3</td>
</tr>
<tr>
<td>TRP</td>
<td>CB</td>
<td>(NE1 + CZ2 + CZ3)/3</td>
</tr>
</tbody>
</table>

Table 7 Definition of anchoring points

For residues with multiple anchoring points, we will normalize the score for standardizing purposes.

4.2. Results and Discussion of OPUS-SSF

We are going to show the OPUS-SSF results in this section. First, OPUS-SSF recognized 571 out of 603 native structures with a Z-score of -5.46. We tested the
scoring function on more decoy sets. The detailed result can be found in Table 8. In testing OPUS-SSF, we want to show the improvement from OPUS-CSF so we added more test sets to test its power. Just like OPUS-CSF, we added a cutoff value of 15 in the calculation. The reason of doing so is similar to OPUS-CSF: a segment with SSF score more than 15 is extremely rare and we want to ignore it. Other results we have are the root mean square deviation (RMSD) values and Pearson’s correlation coefficients between SSF score. The results are improved comparing to the coarse-grained scoring function OPUS-CSF, as we expected since we incorporated the side chain information.

<table>
<thead>
<tr>
<th></th>
<th># of Proteins</th>
<th>OPUS-CSF</th>
<th>OPUS-SSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>7</td>
<td>7 (-3.31)</td>
<td>7 (-5.00)</td>
</tr>
<tr>
<td>fisa</td>
<td>4</td>
<td>2 (-2.55)</td>
<td>2 (-3.84)</td>
</tr>
<tr>
<td>fisa_casp3</td>
<td>5</td>
<td>4 (-6.72)</td>
<td>4 (-13.60)</td>
</tr>
<tr>
<td>hg_structal</td>
<td>29</td>
<td>23 (-2.06)</td>
<td>23 (-2.85)</td>
</tr>
<tr>
<td>ig_structal</td>
<td>61</td>
<td>56 (-2.14)</td>
<td>57 (-3.14)</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>20</td>
<td>20 (-2.08)</td>
<td>20 (-2.84)</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>56</td>
<td>56 (-6.39)</td>
<td>56 (-11.24)</td>
</tr>
<tr>
<td>lattice_ssift</td>
<td>8</td>
<td>8 (-11.79)</td>
<td>8 (-18.92)</td>
</tr>
<tr>
<td>lmds</td>
<td>10</td>
<td>8 (-6.80)</td>
<td>9 (-9.36)</td>
</tr>
<tr>
<td>MOULDER</td>
<td>20</td>
<td>20 (-3.16)</td>
<td>20 (-6.16)</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>58</td>
<td>53 (-4.53)</td>
<td>54 (-5.69)</td>
</tr>
<tr>
<td>casp_good</td>
<td>143</td>
<td>129 (-1.72)</td>
<td>135 (-2.27)</td>
</tr>
<tr>
<td>CASP9</td>
<td>85</td>
<td>46 (-2.80)</td>
<td>84 (-5.91)</td>
</tr>
<tr>
<td>CASP10</td>
<td>43</td>
<td>29 (-4.73)</td>
<td>43 (-8.11)</td>
</tr>
<tr>
<td>CASP11</td>
<td>54</td>
<td>30 (-3.13)</td>
<td>49 (-6.33)</td>
</tr>
<tr>
<td>Total</td>
<td>603</td>
<td>491 (-3.32)</td>
<td>571 (-5.46)</td>
</tr>
</tbody>
</table>

Table 8 The numbers of protein targets in the decoy set, with their native structures successfully recognized by OPUS-CSF and OPUS-SSF are listed in the table. The numbers in parentheses are the average Z-scores of the native structures. The bigger the absolute value of Z-score, the better. Out of totally 603 protein targets in 15 decoy sets, OPUS-SSF recognized 571 native structures with an average Z-score of −5.46, both values are better than that of OPUS-CSF.
Table 9 Average Pearson's Correlation Coefficients of OPUS-CSF and OPUS-SSF Scores with TM-Scores. The correlation coefficient of a decoy set is the average coefficient of all targets in that decoy set. The native structures were excluded from the calculation. OPUS-SSF has better result than OPUS-CSF.

<table>
<thead>
<tr>
<th></th>
<th>OPUS-CSF</th>
<th>OPUS-SSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>-0.67</td>
<td>-0.74</td>
</tr>
<tr>
<td>fisA</td>
<td>-0.55</td>
<td>-0.63</td>
</tr>
<tr>
<td>fisA_casp3</td>
<td>-0.33</td>
<td>-0.36</td>
</tr>
<tr>
<td>hg_structal</td>
<td>-0.80</td>
<td>-0.80</td>
</tr>
<tr>
<td>ig_structal</td>
<td>-0.88</td>
<td>-0.87</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>-0.90</td>
<td>-0.88</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>-0.45</td>
<td>-0.48</td>
</tr>
<tr>
<td>lattice_ssfit</td>
<td>-0.15</td>
<td>-0.08</td>
</tr>
<tr>
<td>lmds</td>
<td>-0.34</td>
<td>-0.32</td>
</tr>
<tr>
<td>MOULDER</td>
<td>-0.86</td>
<td>-0.86</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>-0.39</td>
<td>-0.38</td>
</tr>
<tr>
<td>casp_good</td>
<td>-0.65</td>
<td>-0.60</td>
</tr>
<tr>
<td>CASP9</td>
<td>-0.31</td>
<td>-0.50</td>
</tr>
<tr>
<td>CASP10</td>
<td>-0.25</td>
<td>-0.46</td>
</tr>
<tr>
<td>CASP11</td>
<td>-0.14</td>
<td>-0.29</td>
</tr>
<tr>
<td>Total</td>
<td>-0.52</td>
<td>-0.56</td>
</tr>
</tbody>
</table>
Table 10 Average RMSD Values of OPUS-CSF and OPUS-SSF on 15 Decoys Sets. The numbers for each decoy set are the average RMSD values of recognized structures in that decoy set by two scoring functions. The numbers in the last row are the weighted average numbers of all decoy sets. The value “0” indicates that all native structures were successfully found in that decoy set. OPUS-SSF outperformed OPUS-CSF in every decoy set.

OPUS-SSF is the starting point of combining the side-chain information along with the non-Boltzmann formula. We are going to discuss another implementation of OPUS-Beta in following section.

4.3. Introduction to OPUS-Beta

The success of OPUS-SSF inspired us to develop more scoring functions combining the idea of decomposing the sidechains into rigid body blocks and construct configurational native distributions together to develop more advanced scoring functions a new scoring function named OPUS-Beta, for finding the correct β-sheet contact pattern (the entire residue-residue β-contacts of a protein) among the decoys. OPUS-Beta potential contains six terms, including a main chain carbonyl term, a self-packing term, a pairwise inter-strand packing term, a pairwise intra-strand packing term, a lattice term and a side-chain block contact term. We followed the idea of OPUS-CSF\(^97\) and OPUS-SSF\(^99\) with slight modifications. In OPUS-CSF and OPUS-SSF, they used the entire PDB as training set while in OPUS-Beta, since we are
focusing on beta proteins, for every protein entry, we will use DSSP program to calculate the percentage of beta residues. We only count protein structures with more than 25% residues to remove the influences of non-beta proteins in PDB. We constructed a scoring function based on the distribution of coordinate components of certain anchoring points calculated from short peptide segments in PDB and the distribution of residue contact patterns. An anchoring point is an actual atom position or computed coordinates based on some atoms in the side chain. Similar to OPUS-CSF and OPUS-SSF, we generate the configurational native distribution (CND) look-up table by scanning the filtered PDB in a window of 5, 7, 9, 11 residues. And this look-up table was used for calculating part of Beta score of a test structure with an assumption that every atom coordinate component can be treated as independent Gaussian distribution. In addition to the CND lookup table, we will calculate another lookup table contains the information of beta-residue related information. We assumed the structure with smallest OPUS-Beta score is the native structure.

The performance of OPUS-Beta in recognition of \(\beta\)-sheet topology and residue-residue contacts in \(\beta\)-sheets was tested on the conventional sets for evaluating \(\beta\)-\(\beta\) contact prediction BetaSheet916 and BetaSheet1452 with modification for the recognition process. The results show that OPUS-Beta outperforms other methods in recognition of native \(\beta\)-contact pattern. We expect this new potential to be helpful in improving the prediction of \(\beta\)-contacts in proteins.
4.4. Construction of OPUS-Beta

4.4.1. Construction of lookup table

We followed the basic idea of OPUS-CSF and OPUS-SSF to use the entire PDB as the training set, but we removed non-beta proteins, i.e., it has less than 25% beta residues. To check the β-component of a structure, we used the Dictionary of Protein Secondary Structure (DSSP) \(^{100}\) to assign secondary structures for every residue in the protein. This filtered PDB was used as our training set.

We constructed one configurational native distribution (CND) lookup table of small peptide segments with specific sequences (with length 5, 7, 9, 11). The distributions were calculated through sliding the window length of 5, 7, 9, 11 of every structure in the filtered PDB and the mean atom coordinate and standard deviation is recorded in CND lookup table. In addition to main chain atom coordinate information, we also recorded the side-chain information. Based on the rigid body representation of OPUS-DOSP, we simplified the side-chain information into several anchoring points, the details of finding anchoring points can be found in Table 7 and the details about side blocks can be found in OPUS-DOSP and OPUS-PSP paper. Segments with appearance less than 5 times are discarded from the conclusion of previous work.

The lookup table discussed above is used for calculating main chain carbonyl atom term and side chain block contact term. For other terms, we are constructing another lookup table, with keys as the residue types and corresponding lattice state,
the values are the mean and standard deviation of numbers of such combination. This lookup table can summarize the characteristics of beta contact patterns and help us recognize native beta patterns more efficiently.

4.4.2. Construction of OPUS-Beta score

OPUS-Beta scoring function has six terms: a main-chain carbonyl atom term($E_{main}$), a self-packing term ($E_{self}$), a pairwise inter-strand packing term ($E_{pair\_inter}$), a pairwise intra-strand packing term ($E_{pair\_intra}$), a lattice term ($E_{lattice}$), and a side-chain block contact term($E_{side}$).

$$E = E_{main} + E_{self} + E_{pair\_inter} + E_{pair\_intra} + E_{lattice} + E_{side}$$

**Equation 4-1 Construction of OPUS-Beta score**

Here we did not use weights for every term because we are using entire filtered PDB as training set.

To construct an OPUS-Beta score of a structure, we use a Beta score calculating function. This function works in the following manner: if the input is a single residue, we are calculating the main chain term and side chain packing term. We use the first lookup table to calculate the coordinate component scores of every anchoring point in the side chain and main chain carbonyl atom. If the input is not a single residue, which means we are calculating the Beta sheet related terms. Then we are treating the number of corresponding residues in that state as an independent Gaussian variable and calculate the absolute value of Z- scores using
the second lookup table as our Beta score. For a test structure, we first split this structure into segments of length 5, 7, 9, 11, then for every segment that are in the first CND lookup table, we treat every coordinate component or the residue number as an independent Gaussian variable and calculate the absolute value of Z-score. Then we use the DSSP program to get second structure information and calculate other terms using the second lookup table. The last step is adding them together as the final beta score for a residue. We assume the structure with smallest Beta score as the native structure.

4.4.2.1. Main chain term

A native $\beta$ sheet pattern is also a native structure; hence we can borrow the idea of OPUS-CSF and use the main chain carbonyl atom as an indicator whether the structure is native. For every residue in the structure, we can calculate the corresponding Beta score:

$$E_{main} = Beta(A_i)$$

Equation 4-2 Main chain term

How to calculate the Beta score using the Beta function is defined in previous section.
4.4.2.2. Self-packing term

The self-packing term describes the preference of a residue type on a β-sheet. We specifically took into account different types of the β-strand on which the β-residues connect. The formula for obtaining the self-packing term is:

\[ E_{self}(A_i, L_i) = Beta(A_i, L_i) \]

Equation 4-3 Self-packing term

where \( A_i \) is the residue type at lattice state \( L_i \). The state \( L_i \) in sheet lattice is determined by the topologies of neighbor β-strands(Figure 4-1). Beta denotes the OPUS-Beta calculating function. The process of this function is defined above. Hence, there are a total of 5 states for \( L_i \): one antiparallel partner, one parallel partner, two antiparallel partners, two parallel partners and two partners including one antiparallel partner and one parallel partner. Given that there are 20 different kinds of amino acids, we have \( 20 \times 5 = 100 \) distinguished states for the self-packing energy term.
4.4.2.3. Pairwise Packing term

Presumably, the packing interactions between an inter-strand residue pair could be different from those between an intra-strand residue pair. Thus, we consider two different pairwise packing terms, each of which corresponds to the packing interactions of an inter-strand residue pair and an intra-strand residue pair respectively.

The pairwise packing term for an intra-strand pair and inter-strand pair has the form of

\[ E_{\text{pair, intra}}(A_i, A_j, L_{ij}) = \beta(A_i, A_j, L_{ij}) \]
The difference lies in the $L_{ij}$ state, whether it is an intra-strand state or an inter-strand pair. As the side chains of residues along a β-strand point upwards and downwards alternatively on a β-sheet, only the residue pairs whose side chains have the same direction are able to potentially contact with each other. Therefore, residue pairs $(A_i, A_j)$ are included in the calculation only when their relative positions on β-strands and their side-chain orientation (upwards or downwards) allow sidechain packing interaction. Following the treatment from our previous study 64, we consider all possible inter-strand contacting residue pairs. For the intra-strand pairs, we only consider the next nearest neighbor residues within the same strand. Consequently, the relative position $L_{ij}$ of the two residues in a β-sheet lattice is categorized into seven (four types for antiparallel strands and three types for parallel strands) types for inter-strand partners, and two types for intra-strand partners Figure 4-2.
Figure 4-2 Schematic illustration of different types of pairwise packing patterns between two β-residues. (a) All possible pairwise residue-residue contacts in two antiparallel strands, including four types of inter-strand pairs and two types of intra-strand pairs. AA', a hydrogen-bond-involving inter-strand pair (illustrated by $L_{i-1,j+1}$ based on the diagram); AB', a hydrogen-bond-involving residue interacting with the next hydrogen-bond-involving residue on the opposite strand (illustrated by $L_{i-1,j-1}$ based on the diagram); aa', a non-hydrogen-bond-involving inter-strand pair (illustrated by $L_{i,j}$ based on the diagram); ab', a non-hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand (illustrated by $L_{i,j}$ based on the diagram); ab, a non-hydrogen-bond-involving intra-strand pair (illustrated by $L_{i+2,j}$ based on the diagram); AB, a hydrogen-bond-involving intra-strand pair (illustrated by $L_{i+1,j+1}$ based on the diagram). (b) All possible pairwise residue-residue contacts in two parallel strands, including three types of inter-strand pairs and two types of intra-strand pairs. Aa, a hydrogen-bond-involving residue interacting with a non-hydrogen-bond-involving residue (illustrated by $L_{i-1,j-1}$ based on the diagram); Ab, a hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand toward the C terminus (illustrated by $L_{i-1,j+1}$ based on the diagram); Ba, a hydrogen-bond-involving residue interacting with the next non-hydrogen-bond-involving residue on the opposite strand toward the N terminus (illustrated by $L_{i+1,j-1}$ based on the diagram); ab, a non-hydrogen-bond-involving intra-strand pair
(illustrated by $L_{j-1,j+1}$ based on the diagram); AB, a hydrogen-bond-involving intra-strand pair (illustrated by $L_{i-1,i+1}$ based on the diagram).

4.4.2.4. Lattice term

The lattice term describes the possible four-residue contact interactions within $\beta$-sheets. It involves the four residues in a square unit of the $\beta$-lattice, for example, $(i-1, i+1, j-1, j+1)$ in Figure 4-2. Once again, we mainly focus on the effects of the side chain interaction, so we count only the next nearest neighbors. The lattice term sets a global topology regulation of the $\beta$-sheets. Due to limit number of experimentally solved protein structures, we ignore the order of the different residues for obtaining a balance between the detailed description of $\beta$-sheet contact patterns and the number of possible energy states. The term has a form of

$$E_{lattice}(A_i, A_j, A_k, A_l) = Beta(A_i, A_j, A_k, A_l, L_{ijkl})$$

Equation 4-5 Lattice term

The term $A_i, A_j, A_k, A_l$ are the residue types and $L_{ijkl}$ is the corresponding lattice state.

4.4.2.5. Side Chain Block Contact term

For this term, we followed the previous work of OPUS-PSP and OPUS-DOSP. Again, a native $\beta$-contact pattern should also be a native structure. Since a
combination of side chain packing and main chain carbonyl information OPUS-SSF yielded satisfying performance on native structure recognition, we introduce this term into OPUS-Beta. The basic idea is coarse-graining the sidechain of 20 amino acids into 19 type of blocks and calculate the Beta score using the Beta calculating function.

\[ E_{side} = Beta(A_i) \]

\textbf{Equation 4-6 Side Chain Block Contact term}

\textbf{4.4.3. Results of OPUS-Beta}

To test the performance of the OPUS-Beta potential in recognizing native structures from structures with non-native β-sheet contacts, we used the two conventional sets for evaluating β-β contact prediction Beta916 and Beta1452 with modifications. We downloaded the full structure from PDB. Owing to the rich β contents of the test proteins, the decoy set collection covers a substantially large variation of β-contact patterns, allowing a thorough test on the effectiveness of a β-contact scoring function. We generated the decoys in two ways. First, we generated decoys, for each protein, by shifting a specific β-strand (longer than two residues in length) along the direction of the strand forward or backward up to 3 residues or by completely reversing the direction of that β-strand. Therefore, for a protein with N eligible β-strands, there are \((3 \times 2 + 1)N\) decoys. Decoys generated this way are suitable for tests on the efficiency of a potential function in recognizing the right β registration. For method two, we generated decoys, for each protein, by swapping a
pair of β-strands in a β-sheet. If the two β-strands are of different lengths, we only swap the residues in the shorter one with part of the residues of the longer one near the amino terminus. As a result, we have $N(N - 1)/2$ decoys for each protein. These decoys are suitable for tests on the efficiency of a potential in recognizing the right β-sheet topology.

In the decoy recognition tests, we compared the OPUS-Beta scoring function with results from RDb2C2$^{103}$ and bbcontact$^{104}$. We extracted pair-wise pseudo-energy from these methods and used them to recognize native structures from decoys.

The results for the tests on decoy set are shown in Figure 4-3. For Beta916, OPUS-Beta successfully distinguished 750 out of 916 native β-contact pattern from the decoys, while for Beta1452, OPUS-Beta successfully recognized 1154 out of 1452 native structures from decoys. OPUS-Beta outperformed all methods aforementioned. The robust performance indicates the effectiveness of our method.
Figure 4-3 The performance of different kinds of potentials on decoy set Beta916 and Beta1452. The x axis is different scoring functions. The y axis is the number of successfully recognized native structures from their decoys. The result of RDb2C2 and bbcontact are the result of pseudo-energy derive from the produced contact map.
Chapter 5

Conclusion

In this thesis, we first review the basic concepts of protein structure and protein folding process, specifically, we discussed the computational approach to predict protein structures and a particularly useful tool: empirical potential or statistical potentials. Then, in Chapter 2, we introduced OPUS-DOSP, an all-atom potential with side-chain coarse-grained. OPUS-DOSP potential is a distance-dependent and orient-dependent all-atom statistical potential. We discussed the orientation term, the distance term, and the auxiliary function used in OPUS-DOSP. The orientation term of OPUS-DOSP is based on the work of OPUS-PSP. We briefly reviewed OPUS-PSP. It constructed a basis set by decomposing the side chain structures of 20 amino acid residues into 19 block types. Then the statistics of block contact patterns in a non-redundant database was collected to capture the essential information of sidechain contact patterns and packing orientation information without losing too much conciseness, which is a disadvantage of coarse-grained approaches.
In Chapter 3, we tried to introduce a coarse-grained scoring function OPUS CSF. Unlike traditional empirical potentials that is based on Boltzmann formula, OPUS-CSF is based on the native distributions of coordinate components of mainchain C (carbonyl) atoms on a few selected residues of small peptide segments of 5, 7, 9, and 11 residues in length. We also talked about the scoring function, termed as CSF scoring function, was calculated for a particular test structure by comparing the information of its segments with the CND lookup table.

In the last chapter, two combinations of OPUS-DOSP and OPUS-CSF is discussed. OPUS-SSF improved the native structure recognition a lot, OPUS-Beta is focused on Beta-sheet topology ranking. These two scoring functions combine the side-chain processing techniques applied from OPUS-DOSP and the idea of non-Boltzmann formula in OPUS-CSF. The result is promising and thus may contribute to the protein structure prediction in the future.


67. Yang, Y.; Zhou, Y., Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.* **2008**, *17* (7), 1212-1219.


Appendix A

We almost share the same definition of reference frame as OPUS-PSP with some changes and modifications. The modified tables are shown in Tables A1, A2. In Table A1, we specify the atoms used to build the local reference frame, and in Table A2 we provide the detailed method to build the local reference frame. The details on how to determine the inter-rotation angle $\psi_{ab}$ is shown in Figure A1 and Table A3. It is identical to the method in OPUS-PSP.$^{70}$
Table A1: Atoms used to define the reference frame of each block type. The definitions are basically identical to PSP potential with some modifications. In this table, $\text{C}_\alpha$ refers to the $\text{C}_\alpha$ atom in that residue, $\text{Co}$ refers to the position of $(\text{C}_\alpha + 0)/2$ where 0 means the oxygen atom of peptide in that residue. “Prev.” means the N atom in the residue, and “Next.” means the O atom in the residue. The atom index of each atom is given in Figure 2-1a.

<table>
<thead>
<tr>
<th>Block Class</th>
<th>I</th>
<th>II, III</th>
<th>IV, V, VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\text{O}=r_1$</td>
<td>$\text{O}=\frac{r_1+\text{Co}}{2}$</td>
<td>$\text{O}=\frac{r_1+\text{Co}}{2}$</td>
</tr>
<tr>
<td>$\text{x}$</td>
<td>$\text{x} = \frac{r_2 - r_3}{</td>
<td></td>
<td>r_2 - r_3</td>
</tr>
<tr>
<td>$\text{y}$</td>
<td>--</td>
<td>--</td>
<td>$a = r_5 - r_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\tilde{y} = \frac{a - (\text{Co})\text{x}}{</td>
</tr>
<tr>
<td>$\text{z}$</td>
<td>--</td>
<td>--</td>
<td>$\tilde{z} = \frac{\text{x} \times \tilde{y}}{</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Block Class</th>
<th>VI</th>
<th>VIII</th>
<th>IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\text{O} = \frac{r_2 + r_3}{2}$</td>
<td>$\text{O} = \frac{r_2 + r_3 + r_4 + r_5}{4}$</td>
<td>$\text{O} = \frac{r_2 + r_3 + r_4 + r_5}{4}$</td>
</tr>
<tr>
<td>$\text{x}$</td>
<td>$\text{x} = \frac{(r_2 - r_3) \times (r_2 - r_3)}{</td>
<td></td>
<td>(r_2 - r_3) \times (r_2 - r_3)</td>
</tr>
<tr>
<td></td>
<td>$\text{y} = \frac{a - (\text{Co})\text{x}}{</td>
<td></td>
<td>a - (\text{Co})\text{x}</td>
</tr>
<tr>
<td>$\text{z}$</td>
<td>--</td>
<td>$\tilde{z} = \frac{\text{x} \times \tilde{y}}{</td>
<td></td>
</tr>
</tbody>
</table>

Table A2: Definition of the local reference frame of each block class.
To determine the inter-rotation angle $c$, we consider two hypothetical planes, $P_a$ and $P_b$, that are perpendicular to the relative direction $r_{a \rightarrow b}$ (i.e., $P_a \parallel P_b$). The vectors $u_{xa}$ and $u_{ya}$ in plane $P_a$ (for block type a) and $u_{xb}$ and $u_{yb}$ in $P_b$ (for block type b) are determined from Table A3b and then applied to Table A3c. In this example, we find that $u_{ya}$ is roughly parallel to $u_{xb}$ and pointing in the opposite direction, so we classify the relationship as $u_{ya} \parallel -u_{xb}$. Similarly, we classify $u_{ya}$ and $u_{ya}$ as $u_{xa} \parallel u_{yb}$.

Figure A1 Determining coarse-grained inter-rotation angle $\psi_{ab}$
<table>
<thead>
<tr>
<th>( \vec{r}_{a \rightarrow b} : (n_x, n_y, n_z) )</th>
<th>Symmetry</th>
<th>( \mathbf{u}_x )</th>
<th>( \mathbf{u}_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_y = n_z = 0 )</td>
<td>( A_z )</td>
<td>( \hat{z} )</td>
<td>( \hat{y} )</td>
</tr>
<tr>
<td>( n_z \neq 0 )</td>
<td>Others</td>
<td>( \hat{y} )</td>
<td>( \hat{z} )</td>
</tr>
<tr>
<td>( n_z = n_x = 0 )</td>
<td>( A_x )</td>
<td>( \hat{x} )</td>
<td>( \hat{z} )</td>
</tr>
<tr>
<td>( n_y \neq 0 )</td>
<td>Others</td>
<td>( \hat{z} )</td>
<td>( \hat{x} )</td>
</tr>
<tr>
<td>( n_x = n_y = 0 )</td>
<td>( A_y )</td>
<td>( \hat{y} )</td>
<td>( \hat{x} )</td>
</tr>
<tr>
<td>( n_z \neq 0 )</td>
<td>Others</td>
<td>( \hat{x} )</td>
<td>( \hat{y} )</td>
</tr>
<tr>
<td>( n_x = 0 )</td>
<td>Any</td>
<td>( \hat{x} )</td>
<td>( \frac{n_y \hat{y} - n_z \hat{z}}{\sqrt{2}} )</td>
</tr>
<tr>
<td>( n_y n_z \neq 0 )</td>
<td>Any</td>
<td>( \hat{y} )</td>
<td>( \frac{n_z \hat{z} - n_x \hat{x}}{\sqrt{2}} )</td>
</tr>
<tr>
<td>( n_y = 0 )</td>
<td>Any</td>
<td>( \hat{y} )</td>
<td>( \frac{n_x \hat{x} - n_y \hat{y}}{\sqrt{2}} )</td>
</tr>
<tr>
<td>( n_z n_x \neq 0 )</td>
<td>Any</td>
<td>( \hat{z} )</td>
<td>( \frac{2n_z \hat{z} - n_x \hat{x} - n_y \hat{y}}{\sqrt{6}} )</td>
</tr>
<tr>
<td>( n_x = 0 )</td>
<td>Any</td>
<td>( \frac{n_x \hat{x} - n_y \hat{y}}{\sqrt{2}} )</td>
<td>( \frac{n_y \hat{y} - n_z \hat{z}}{\sqrt{2}} )</td>
</tr>
</tbody>
</table>
Table A3: Definition of inter-rotation bins for pairs of blocks. (a) Nine pairwise inter-rotation symmetry cases, denoted as \(\uparrow\), \(\uparrow\uparrow\), H, \(\uparrow\ A\), \(\uparrow\ P\), AA, AP, PP, and \(PP_{same}\). (b) Definitions of the axes \(u_x\) and \(u_y\) used to determine the inter-rotation bin, as illustrated in Figure A1. The reference frame axes \(\hat{x}\), \(\hat{y}\), and \(\hat{z}\) are derived from Table A1. (c) Criteria for determining the inter-rotation bin \(\psi_{ab}\). The properties \(u_a||u_b\) and \(u_a \perp u_b\) are illustrated in Figure A1. Note that “|” represents any line.
shape local symmetry (e.g., ↓, ↑) and “?” represents the local symmetry of any shape.

Furthermore, we use “F” to emphasize the difference between PP and same $PP^{same}$.
Appendix B

The reason of not using a continuous auxiliary function are as follows: we think the noise in our result may reverse the sign of our potential, especially the values close to 0. Therefore, we designed a function which is not continuous at 1 to create an energy gap, and this gap will make it much harder to reverse the sign of our result. To justify our assumption, we tested the result for another version of OPUS-DOSP, which we used a function that is continuous at 1.

\[
g(x) = \begin{cases} 
1 - x & x > 1 \\
0 & x = 1 \\
1 & x < 1 \\
\end{cases}
\]

Where \( x = \frac{p_{obs}}{p_{ref}} \)

The result is shown in Table B1, and it shows the necessity of adding a gap in our energy function.

<table>
<thead>
<tr>
<th>Decoy sets</th>
<th>Total numbers of targets</th>
<th>DOSP</th>
<th>DOSP (without energy gap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4state_reduced</td>
<td>7</td>
<td>3(-4.18)</td>
<td>3(-3.74)</td>
</tr>
<tr>
<td>fisa</td>
<td>4</td>
<td>2(-4.73)</td>
<td>2(-3.17)</td>
</tr>
<tr>
<td>fisa_casp3</td>
<td>5</td>
<td>4(-4.33)</td>
<td>4(-4.31)</td>
</tr>
<tr>
<td>hg_structal</td>
<td>29</td>
<td>27(-3.57)</td>
<td>27(-3.40)</td>
</tr>
<tr>
<td>ig_structal</td>
<td>61</td>
<td>61(-7.15)</td>
<td>61(-7.08)</td>
</tr>
<tr>
<td>ig_structal_hires</td>
<td>20</td>
<td>20(-4.24)</td>
<td>20(-4.23)</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>56</td>
<td>51(-5.05)</td>
<td>46(-4.33)</td>
</tr>
<tr>
<td>lattice_sxfit</td>
<td>8</td>
<td>3(-4.46)</td>
<td>3(-4.14)</td>
</tr>
<tr>
<td>lmds</td>
<td>10</td>
<td>10(-7.68)</td>
<td>10(-7.39)</td>
</tr>
<tr>
<td>MOULDER</td>
<td>20</td>
<td>17(-4.12)</td>
<td>19(-4.58)</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>58</td>
<td>51(-3.92)</td>
<td>52(-4.35)</td>
</tr>
<tr>
<td>Total(Z-score)</td>
<td>278</td>
<td>249(-5.02)</td>
<td>247(-4.02)</td>
</tr>
</tbody>
</table>
Table B1. The performance of OPUS-DOSP with and without the energy gap in energy function on 11 decoy sets. The numbers in parentheses are the average Z-scores of the native structures. The performance with a gap in energy function (column: DOSP) is better than the performance without the gap (column: DOSP (without energy gap))

We set up an energy cutoff $s$ in OPUS-DOSP because we find out that if we add the auxiliary function, sometimes, because of some individual extreme values, the potential may be extremely large and unreliable. We set up a cutoff $s$ to keep the value of $\frac{p_{obs}}{p_{ref}}$ between $1/s$ and $s$. If the ratio is lower than $c$, we set the ratio to $1/s$. If the ratio is bigger than $s$, we set the ratio to $s$. Accordingly, the potential value is set to $s$ and $-s$. The specific value of $s$ can be obtained by our logarithm result. We list some quantile values of the logarithm absolute energy in Table B2. From the logarithm result, we find out that 99% of the absolute value of energy is below 4, so we set the cutoff values as $e^4 \approx 55$, that is a reasonable value of cutoff energy. What is more, based on the idea of emphasizing extreme cases, we add the cutoff value $s$ to cases that are missing in training result, which is the bias favoring existing structures, or the penalty towards cases not found in native structures. To test our assumption, we calculate the result when we do not set cutoff on potential and the result when we do not add bias to missing structures. In Figure B1, we can conclude that the performance becomes worse if we do not add the cut off $s$, or we don't add
bias, which means that the idea of setting up a cut-off energy and penalizing missing structures is reasonable.

Table B2 Some quantile values of absolute logarithm energy. “99%” means 99% of absolute logarithm energy is smaller than this value, i.e., 3.999.
Figure B1. The performance of OPUS-DOSP without energy-cutoff and bias penalizing missing structures. (a) is the success number and (b) is the average Z-score.