JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2019, 15, 5154−5160

pubs.acs.org/JCTC

OPUS-Rota2: An Improved Fast and Accurate Side-Chain Modeling Method

Gang Xu,[†,‡,∇] Tianqi Ma,[§,∥,∇] Junqing Du,[⊥] Qinghua Wang,[⊥] and Jianpeng Ma*[,†,‡,§,∥,⊥,#]

[†]Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 200433, China

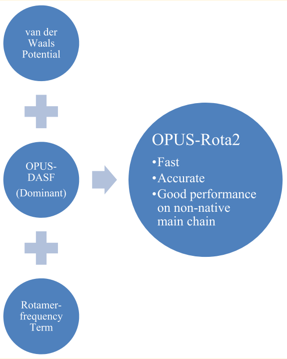
[‡]School of Life Sciences, Tsinghua University, Beijing 100084, China

[§]Applied Physics Program and [∥]Department of Bioengineering, Rice University, Houston, Texas 77005, United States

[⊥]Verna and Marrs Mclean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, BCM-125, Houston, Texas 77030, United States

[#]School of Life Sciences, Fudan University, Shanghai 200433, China

ABSTRACT: Side-chain modeling plays a critical role in protein structure prediction. However, in many current methods, balancing the speed and accuracy is still challenging. In this paper, on the basis of our previous work OPUS-Rota (*Protein Sci.* **2008**, *17*, 1576−1585), we introduce a new side-chain modeling method, OPUS-Rota2, which is tested on both a 65-protein test set (DB65) in the OPUS-Rota paper and a 379-protein test set (DB379) in the SCWRL4 paper. If the main chain is native, OPUS-Rota2 is more accurate than OPUS-Rota, SCWRL4, and OSCAR-star but slightly less accurate than OSCAR-o. Also, if the main chain is non-native, OPUS-Rota2 is more accurate than any other method. Moreover, OPUS-Rota2 is significantly faster than any other method, in particular, 2 orders of magnitude faster than OSCAR-o. Thus, the combination of higher accuracy and speed of OPUS-Rota2 in modeling side chains on both the native and non-native main chains makes OPUS-Rota2 a very useful tool in protein structure modeling.



■ INTRODUCTION

Protein structure prediction has become increasingly important and powerful over the past few decades. However, determining a protein structure exclusively from an amino acid sequence is still very challenging.[1] One way to improve the performance of protein structure prediction is to develop a fast and accurate side-chain modeling method, which is also essential for refining the high-accuracy structure models. This side-chain modeling method should satisfy two requirements: First, it should be fast enough for the time efficiency. Second, it should be accurate for cases in which the main chains are in either native or non-native states.

To minimize the sampling space, most side-chain modeling methods explore a limited number of representative conformations in the rotamer library that are derived from a set of high-resolution X-ray structures.[2−8] In these methods, some of them enhance the sampling scheme to sample the rotamers with more efficiency,[9−24] some of them improve the empirical potential functions to obtain better optimization,[21,25−37] and some of them alter the standard rotamers slightly to increase the diversity of the conformation.[38−40] Also, there are some nonrotameric methods.[41,42]

In side-chain modeling, it is hard to balance the accuracy and the speed. Some methods, such as SCWRL4,[37] sacrifice their accuracy by using a simplified pairwise energy function and dead-end elimination[9] to achieve a relatively higher speed. In contrast, some methods such as NCN,[30] LGA,[27] and OSCAR,[35,36] achieve better accuracy by using a more complicated and accurate potential function,[21,25] which is computationally more expensive. To leverage the performance between accuracy and speed, our previous work OPUS-Rota[43] uses a simple yet effective potential OPUS-PSP[44] in its energy function, which reduces the computational burden and achieves good accuracy. After OPUS-PSP, many more empirical potentials have been developed, such as OPUS-DOSP[45] and OPUS-CSF.[46] These more powerful potentials make it possible to develop a more accurate and faster side-chain modeling method.

In this paper, we develop a new side-chain modeling method, OPUS-Rota2, which is based on our previous work OPUS-Rota.[43] We examine the performance of OPUS-Rota2 on a 65-protein benchmark set (DB65) used in OPUS-Rota[43] and a 379-protein benchmark set (DB379) used in SCWRL4.[37] If the main chain is native, OPUS-Rota2 is more accurate than OPUS-Rota, SCWRL4, and OSCAR-star but slightly less accurate than OSCAR-o. Also, if the main chain is non-native, OPUS-Rota2 is more accurate than any other method. OPUS-Rota2 is significantly faster than any other method, in particular, 2 orders of magnitude faster than OSCAR-o. The high accuracy and computational efficiency make OPUS-Rota2 a very useful side-chain modeling tool in the early stage of protein structural modeling.

Received: March 29, 2019
Published: August 14, 2019

DOI: 10.1021/acs.jctc.9b00309

■ METHODS

In OPUS-Rota2, we sample the side-chain rotamer from a backbone-dependent rotamer library[47] for each residue in a random order. The process also involves an energy function that contains multiple terms. The final side-chain structures are reconstructed by minimizing the energy function. A distinct feature of this work is that its energy function contains a special scoring function term called the OPUS-DASF term that describes relative positions of atoms on the side chains. The OPUS-DASF term is used as a replacement for the OPUS-PSP term[44] in the OPUS-Rota method[43] developed by us earlier.

Energy Function. The total energy function contains four terms

$$E_{\text{total}} = w_{\text{dasf}}E_{\text{dasf}} + w_{\text{vdw\_mc}}E_{\text{vdw\_mc}} + w_{\text{vdw\_sc}}E_{\text{vdw\_sc}} + w_{\text{rot}}E_{\text{rot}}$$

where E_{dasf} is the OPUS-DASF term (defined below), $E_{\text{vdw\_mc}}$ is the modified LJ potential for an atom pair between a main-chain atom and a side-chain atom, $E_{\text{vdw\_mc}}$ is the modified LJ potential for an atom pair between two side-chain atoms that are on different residues, and E_{rot} is a term related to rotamer frequency. Since the main chain is fixed, the LJ potential between different main-chain atoms is not included. Just as in OPUS-Rota, the four weights $w_{\text{dasf}} = 0.1$, $w_{\text{vdw\_mc}} = 0.8$, $w_{\text{vdw\_sc}} = 0.6$, and $w_{\text{rot}} = 1.0$ are optimized against a small set of high-resolution structures containing 1aac, 1bpi, 1isu, 1ptx, 1xn, 256b, 2erl, 2hbg, 2ihl, 5rxn, and 9rnt.

Compared to the energy function in OPUS-Rota, we use the OPUS-DASF term to replace the OPUS-PSP term. In addition, we also remove the solvation energy term for time efficiency.

OPUS-DASF Term. We propose a dihedral angle scoring function (DASF), which is based on our previous work OPUS-CSF.[46] OPUS-CSF is a fast and accurate scoring function that can be used to distinguish the native protein structures from their decoy base on main-chain configurations.

Similar to OPUS-CSF, we generate four DASF lookup tables for small peptide segments of 5, 7, 9, and 11 residues in length, respectively, by scanning through the entire Protein Data Bank (PDB). In order to model the side-chain conformation, we save the coordinate of the side-chain atoms that would be used to calculate the side-chain dihedral angles. The identities of the recorded atoms are listed in Table 1. The segments of different lengths are marked as 5(1, 3, 5), 7(2, 4, 6), 9(1, 3, 5, 7, 9), and 11(2, 4, 6, 8, 10). In the form of 5(1, 3, 5), for example, the first number 5 is the segment length; the numbers 1, 3, 5 in the parentheses are the indices of the recorded residues whose coordinates of recorded atoms are saved.

Different from OPUS-CSF, in the case of side-chain modeling, since the main chain is fixed, we can exclude the influence of main-chain atoms. Therefore, we build the local molecular coordinate system on each recorded residue separately, instead of building one on the center residue as described in OPUS-CSF. The local molecular coordinate system is constructed via the coordinates of the main-chain C atom, Ca atom, and O atom. The Ca atom is set as the origin, and the line connecting the Ca and C atoms is defined as the *X*-axis. The parallel component of the C−O vector that is perpendicular to the *X*-axis in the Ca−C−O plane is defined as the *Y*-axis, and the *Z*-axis is defined correspondingly. More details can be found in the OPUS-CSF paper.[46]

Table 1. Recorded Atoms in 20 Different Residues[a]

	F1	F2	F3	F4
GLY				
ALA				
SER	OG			
CYS	SG			
VAL	CG1			
ILE	CG1	CD		
LEU	CG	CD1		
THR	OG1			
ARG	CG	CD	NE	CZ
LYS	CG	CD	CE	NZ
ASP	CG	OD1		
GLU	CG	CD	OE1	
ASN	CG	OD1		
GLN	CG	CD	OE1	
MET	CG	SD	CE	
HIS	CG	ND1		
PRO	CG	CD		
PHE	CG	CD1		
TYR	CG	CD1		
TRP	CG	CD1		

[a] The recorded atoms will be used to calculate the side-chain dihedral angles. It should be noticed that different residues may have different numbers of recorded atoms.

The DASF lookup table is constructed as follows: first, we assume that the coordinate components of a specific recorded atom are independent from the same recorded atom in other segments that have the same sequence in the PDB, and they obey a Gaussian distribution. Then, after scanning through the entire PDB, we gather the coordinate components of each recorded atom with same the segment sequence and use them to generate the Gaussian distributions for each recorded atom, respectively. Finally, we calculate the means and standard deviations of the distributions and save them in the DASF lookup table. We construct four DASF lookup tables depending on four different segment lengths.

For protein structure evaluation, we first divide the structure by all possible overlapping segments (5, 7, 9, 11 residues in length). Then, for every segment found in the corresponding DASF lookup table, we calculate the absolute values of the *Z*-score for each coordinate component of the recorded atoms on the basis of the distributions in the DASF lookup table. In the end, the absolute *Z*-scores of all coordinate components of the recorded atoms are added up to form the DASF score. We use Sn to denote the DASF score for the *n*-residue segment, and it is calculated as follows:

$$\text{Sn} = \sum_{n\text{-residuesegment}} \sum_{\text{recordedatomsinresidue}} \left(\left| \frac{X_{\text{test}} - X_{\text{mean}}}{\delta_x} \right| + \left| \frac{Y_{\text{test}} - Y_{\text{mean}}}{\delta_y} \right| + \left| \frac{Z_{\text{test}} - Z_{\text{mean}}}{\delta_z} \right| \right)$$

Here, $n \in [5, 7, 9, 11]$, $(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}},)$, are the coordinate components of a recorded atom in its local molecular coordinate system; $(X_{\text{mean}}, Y_{\text{mean}}, Z_{\text{mean}},)$ and $(\delta_x, \delta_y, \delta_z,)$ are the means and standard deviations of the coordinate components of the recorded atom in the DASF look up table.

In OPUS-DASF, we suppose that the longer peptide segments contain more conservative and reliable information.

The final DASF score is constructed using different weights for different segment lengths

$$E_{\text{dasf}} = \sum_{n} \text{Sn} \times \frac{n}{10}$$

Here, $n \in [5, 7, 9, 11]$.

van der Waals (vdW) Potential. Just as in OPUS-Rota,[43] the van der Waals potential between two atoms i and j takes the form

$$E_{\text{vdW}}(i, j) = \begin{cases} \lambda(49.69 - 40.06d_{ij}^{*}),\ d_{ij}^{*} \in [0,\ 1/1.33] \\ \lambda e_{ij}((d_{ij}^{*})^{-12} - 2(d_{ij}^{*})^{-6}),\ d_{ij}^{*} \in [1/1.33,\ 1/1.12] \\ e_{ij}((d_{ij}^{*})^{-12} - 2(d_{ij}^{*})^{-6}),\ d_{ij}^{*} \in [1/1.12,\ 2.5] \end{cases}$$

where $d_{ij}{}^{*} = d_{ij}/a_{ij}$, and d_{ij} is the distance between atoms i and j. $a_{ij} = a_i + a_j$, with a_i and a_j as the atomic radii.$e_{ij} = \sqrt{e_i e_j}$, and e_i and e_j are the well depths. The constant λ is the scaling factor for the repulsive term, which is set to unity if both atoms i and j are aromatic carbons and is 1.6 otherwise. The details of the LJ parameters and summation rules are described in our OPUS-PSP paper.[44] To speed up the calculation, the LJ cutoff distance is $d_{ij}{}^{*} = 2.5$.

In a real application, we consider $E_{\text{vdw\_mc}}$ and $E_{\text{vdw\_sc}}$ separately.

Rotamer-Frequency-Related Term. The rotamer-frequency-related term is defined as follows:

$$E_{\text{rot}} = -\sum_{m=1}^{N} r \log \frac{p(R_m | \varnothing_m,\ \varphi_m,\ A_m)}{p(R_m = 1 | \varnothing_m,\ \varphi_m,\ A_m)}$$

$$r = \begin{cases} 0,\ A_{m \in \{\text{Gly,Ala}\}} \\ 1,\ \text{otherwise} \end{cases}$$

In this equation, N is the number of residues, and m is the residue index where $m \in [1, N]$. $p(R_m | \varphi_m,\ \phi_m,\ A_m)$ is the occurrence probability of a rotamer R_m whose main-chain torsional angles are φ_m and ϕ_m and residue type is A_m. $p(R_{m=1} | \varphi_m,\ \phi_m,\ A_m)$ is the occurrence probability of the most likely rotamer ($R_{m=1}$). The rotamer frequency is derived from Dunbrack's rotamer library.[47] Unlike for OPUS-Rota, we set r to zero for Gly and Ala, and to unity otherwise.

In OPUS-Rota2, a cutoff value in the rotamer-frequency-related term is used to provide a chance for the rotamers that have very low occurrence probability in the rotamer library. In this case, if $\log \frac{p(R_m | \varnothing_m, \varphi_m, A_m)}{p(R_{m=1} | \varnothing_m, \varphi_m, A_m)} < -5$, we set it to -5, and if $\log \frac{p(R_m | \varnothing_m, \varphi_m, A_m)}{p(R_{m=1} | \varnothing_m, \varphi_m, A_m)} > 5$, we set it to 5.

Procedures of OPUS-Rota2 Implementation. *Initialization.* The energy function in OPUS-Rota2 contains four terms: a DASF term, two vdW terms with one for an atom pair between a main-chain atom and a side-chain atom and another for an atom pair between two side-chain atoms that on different residues, and last a rotamer-frequency-related term. Since the main chain is fixed, the side-chain configuration is limited because of its finite number of rotamers in the rotamer library. For time efficiency, before sampling, we can calculate and store the energy terms of each rotamer that are independent of the configuration of other residues to avoid repetition in calculation during sampling.

In our energy function, the rotamer-frequency-related term is independent of the configuration of other residues. The DASF term only depends on the type of nearby residues, instead of their positions. For vdW interactions, we divide them into three types: the first type is the one between main-chain atoms on different residues, and this type of interaction can be ignored because the main chain is fixed during side-chain modeling. The second type is the vdW interaction between the main-chain atom and the side-chain atom, and this type of vdW potential can be calculated and stored before sampling because of the fixed main chain and the finite side-chain rotamers. The third type is the vdW interaction between side-chain atoms on different residues, and this type of vdW interaction needs to be calculated during sampling because it depends on the side-chain configurations of other residues. In summary, before sampling, we calculate and store the DASF term, the rotamer-frequency-related term, and the vdW term between main-chain atoms and side-chain atoms of all rotamers in rotamer libraray for all residues.

After initialization, we choose the rotamer with the minimum score of the three terms mentioned above for each residue and reconstruct the side-chain configuration. This reconstructed configuration after initialization is called OPUS-Rota2i.

Neighbor List. In order to accelerate the calculation for the vdW interactions between side-chain atoms on different residues, the same as OPUS-Rota, a neighbor list is built in OPUS-Rota2 before sampling using the inverse triangle inequality: $\|x - y - z\| \geq \|x\| - \|z\|$. Since the main-chain atoms are fixed, C_β can be calculated directly;[48] let $d(i, j)$ be the distance between atoms i and j and $C_\beta(i)$ be the value for the C_β atom corresponding to the residue of atom i. Therefore, $d(i, j) \geq d(C_\beta(i), C_\beta(j)) - \max(d(i, C_\beta(i)) - \max(d(j, C_\beta(j))$, where $\max(d(\cdot))$ is calculated using the configuration reconstructed by all possible rotamers. Given that the distance between any two C_β atoms (C_α atoms for Gly) from different residues is a constant for a fixed main chain, the minimum distance between any two residues can be estimated before the sampling. Therefore, we only need to consider the vdW interactions for the atom pairs between the residue pairs whose vdW cutoff distance is less than 2.5.

Sampling. In OPUS-Rota2, a simplified sampling method is employed. First, the rotamer of each residue in OPUS-Rota2i is used as the starting point of the sampling procedure. Then, 200 rounds of sampling are performed. In each round, we choose each residue in a random order and sample the rotamer from the rotamer library for the chosen residue. After each sampling, we reconstruct the chosen residue using the sampled rotamer and calculate the energy via the complete energy function which includes four terms. If the energy is decreased, we accept the sampled rotamer; otherwise, we keep the old one.

Although the sampling procedure is simpler in OPUS-Rota2 than that in OPUS-Rota for time efficiency, the results show that this simplified sampling method works well.

Data Preparation. *Training Sets.* To generate four DASF lookup tables, we downloaded the entire PDB which contains 150 742 structures from ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb on May 10, 2019. We removed the structures in decoy sets and side-chain modeling test sets.

Native Test Sets. We used both the 65-protein test set (DB65) in the OPUS-Rota[43] paper and the 379-protein test set (DB379) in the SCWRL4[37] paper to evaluate different

Table 2. Results of OPUS-DASF on Five Decoy Sets Compared with GOAP, OPUS-CSF, and OPUS-SSF Results[a]

	total	GOAP	OPUS-CSF	OPUS-SSF	OPUS-DASF
3DRobot	200	94 (−1.86)	189 (−4.86)	186 (−5.24)	183 (−5.55)
Rosetta (3DR)	58	37 (−2.16)	51 (−3.83)	53 (−3.98)	52 (−3.95)
I-Tasser (3DR)	56	18 (−1.66)	36 (−3.47)	38 (−3.81)	38 (−3.40)
Rosetta	58	45 (−3.39)	47 (−5.43)	52 (−5.81)	47 (−4.46)
I-Tasser	56	45 (−4.99)	47 (−7.70)	50 (−9.11)	49 (−8.34)

[a] 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 the *Z*-score is, the better our results are. 3DRobot datasets,[52] which include 3DRobot, Rosetta (3DR), and I-Tasser (3DR), are downloaded from https://zhanglab.ccmb.med.umich.edu/3DRobot/decoys. Rosetta and I-Tasser are the original classical benchmarks.[53,54]

Table 3. Performance of Different Side-Chain Modeling Methods[a]

DB65	SCWRL4	OPUS-Rota	OSCAR-star	OSCAR-o	OPUS-Rota2i	OPUS-Rota2
χ_1	0.86	0.88	0.89	0.90	0.89	0.90
χ_{1+2}	0.68	0.71	0.72	0.74	0.74	0.74
χ_{1+2+3}	0.41	0.46	0.47	0.51	0.50	0.51
$\chi_{1+2+3+4}$	0.30	0.36	0.35	0.39	0.38	0.38
aRMSD	0.67	0.59	0.58	0.53	0.55	0.54
SD	0.58	0.55	0.53	0.53	0.53	0.51
time	1.91	2.98	6.55	583.00	0.55	0.98
DB379	SCWRL4	OPUS-Rota	OSCAR-star	OSCAR-o	OPUS-Rota2i	OPUS-Rota2
χ_1	0.85	0.86	0.87	0.88	0.86	0.87
χ_{1+2}	0.68	0.69	0.70	0.73	0.69	0.71
χ_{1+2+3}	0.39	0.41	0.42	0.47	0.43	0.45
$\chi_{1+2+3+4}$	0.29	0.31	0.32	0.37	0.32	0.33
aRMSD	0.68	0.64	0.62	0.57	0.63	0.61
SD	0.60	0.59	0.57	0.57	0.59	0.58
time	2.13	3.08	6.35	569.05	0.54	0.95

[a] The accuracy of χ_1 is defined as the percentage of residues whose predicted χ_1 is no more than 40° from the native value, and the accuracy of χ_{1+2} is defined as the percentage of residues for which both χ_1 and χ_2 are in the 40° range compared to the native value. The same goes for the χ_{1+2+3} and the χ_{1+2+4} rows. For these indicators, the larger the better. The aRMSD row stands for the average RMSD between all atoms in the predicted residue and the native residue; the smaller its value is, the better. SD stands for the standard deviation of all RMSDs. We ignore the comparison for the residue if any side-chain atom in its native configuration is missing. Time stands for the average time for modeling each structure.

side-chain modeling methods. We cleaned the PDB files in these two test sets with the following rules: First, if a protein has more than one chain, we only use the first chain. Second, if the atom has more than one conformation, we only use conformation A. Third, if any atom in the main chain is missing, we exclude this protein. We separate each clean set into two subsets, one of which contains main-chain atoms only and the other contains all atoms. The two sets can be downloaded from our Web site.

Non-Native Test Sets. To evaluate the performance of different side-chain modeling methods when the configurations of the main chains are non-native, we added random noise with various strengths to the main-chain torsional angles. First, the proteins in DB65 and DB379 are combined to a new test set, DB437. Then, a non-native main-chain test set is constructed using the main-chain torsional angles with their original values multiplied by a modulating factor randomly sampled from a Gaussian distribution for all proteins in DB437. We used 10 different levels of noise strength; i.e., the mean values of Gaussian are 1.0, and the standard deviations of them are (0.001, 0.003, 0.005, 0.008, 0.01, 0.013, 0.014, 0.015, 0.016, 0.02). Thus, we have 10 non-native test sets, with each of them containing 437 proteins. The corresponding average main-chain RMSD[49] values of 437 proteins (between the randomized structure and the native structure) at each noise level are (0.21, 0.57, 0.93, 1.48, 1.88, 2.38, 2.55, 2.74, 2.95, 3.68) Å. All the non-native main-chain test sets we construct can also be download from our Web site.

RESULTS

Performance of OPUS-DASF. We evaluate the performance of OPUS-DASF on five decoy sets. For comparison, the performances of GOAP,[50] OPUS-CSF,[46] and OPUS-SSF[51] are also listed in Table 2. The results show that OPUS-DASF, which used side-chain information exclusively, achieves a comparable performance with OPUS-CSF, which is only based on the main-chain structure. OPUS-SSF combines the information on the side chain and the main chain and delivers the best results.

Performance of OPUS-Rota2 on Native Test Sets. The performance of OPUS-Rota2 is compared with that of SCWRL4,[37] OPUS-Rota,[37] OSCAR-star,[35] and OSCAR-o[36] on test sets DB65 and DB379. The results are shown in Table 3. In terms of the accuracy of χ_1 and χ_{1+2} and the value of aRMSD, OPUS-Rota2 is better than SCWRL4, OPUS-Rota, and OSCAR-star, but slightly worse than OSCAR-o. However, OPUS-Rota2 is significantly faster than any other method, 2× faster than SCWRL4, 3× faster than Rota, and 6× and 500× faster than OSCAR-star and OSCAR-o, respectively. Moreover, the nonsampling version OPUS-Rota2i is 2× faster than OPUS-Rota2 while achieving a comparable accuracy with OSCAR-star.

Contributions of Different Terms in OPUS-Rota2 Energy Function. In order to find out the dominant contributing energy term in OPUS-Rota2, the vdW term, the rotamer-frequency-related term, and the DASF term are used alone to examine the performance. As shown in Table 4, when using the DASF term exclusively, the accuracies of χ_1, χ_{1+2}, χ_{1+2+3}, and $\chi_{1+2+3+4}$ and the value of aRMSD are significantly better than those using the other two terms. These results demonstrate the dominant contribution of the DASF term in the prediction accuracy. In contrast, in OPUS-Rota,[43] the dominant term is the vdW term. Given that the main difference between OPUS-Rota and OPUS-Rota2 is that OPUS-Rota2 uses the OPUS-DASF term while OPUS-Rota uses the OPUS-PSP term, a reasonable assumption is that the improvement of empirical potential leads to the improvement of the side-chain modeling method.

Table 4. Contributions of Different Energy Terms[a]

DB65	vdW term	rotamer term	DASF term	OPUS-Rota2i	OPUS-Rota2
χ_1	0.74	0.73	0.85	0.89	0.90
χ_{1+2}	0.47	0.52	0.66	0.74	0.74
χ_{1+2+3}	0.19	0.24	0.44	0.50	0.51
$\chi_{1+2+3+4}$	0.10	0.16	0.32	0.38	0.38
aRMSD	0.97	0.94	0.66	0.55	0.54
DB379	vdW term	rotamer term	DASF term	OPUS-Rota2i	OPUS-Rota2
χ_1	0.72	0.73	0.79	0.86	0.87
χ_{1+2}	0.44	0.53	0.57	0.69	0.71
χ_{1+2+3}	0.17	0.24	0.35	0.43	0.45
$\chi_{1+2+3+4}$	0.08	0.17	0.25	0.32	0.33
aRMSD	1.03	0.94	0.81	0.63	0.61

[a] The energy function of each version of OPUS-Rota2 contains only one corresponding energy term.

Performance of OPUS-Rota2 on Non-Native Test Sets. To evaluate the performance of different side-chain modeling methods on non-native main chains, we compare the performances of OPUS-Rota2i and OPUS-Rota2 with those of SCWRL4, OSCAR-star, OSCAR-o, and OPUS-Rota on three non-native test sets. The main-chain aRMSD values between non-native and original native structures in these three non-native test sets are 0.93, 1.88, and 3.68 Å, respectively. Figure 1 shows that, in terms of χ_1 accuracy, both OPUS-Rota2i and OPUS-Rota2 deliver better performance than other methods in all three non-native test sets. With the same modeling method, as the main-chain aRMSD increases, the performance decreases somewhat as expected; i.e., for all methods, the blue bar is the tallest and the green bar is the shortest. For the relative decrease between the green bars vs the blue bars, both OPUS-Rota2i and OPUS-Rota2 methods have the smallest rangeability compared with other methods, indicating that both OPUS-Rota2i and OPUS-Rota2 methods are more resilient to the increasing deviations of the main-chain conformations.

We further examine the performance of OPUS-Rota2i and OPUS-Rota2 on 10 non-native main-chain test sets, and the results are shown in Figure 2. It seems that OPUS-Rota2 is better than OPUS-Rota2i when the main chains are very close to the native ones (<0.7 Å), while OPUS-Rota2i outperforms OPUS-Rota2 as the main-chain aRMSD increases. This is an interesting feature as OPUS-Rota2i is faster than OPUS-Rota2; thus, it may be advantageous to use OPUS-Rota2i at the earlier stages with a lower quality of main chains, and OPUS-Rota2 should be used when the main chains are closer to the native states.

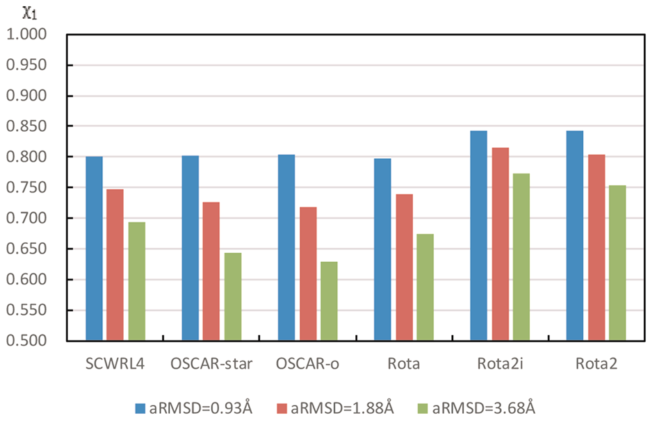


Figure 1. Performance of different side-chain modeling methods on different non-native test sets. Three non-native test sets (each containing 437 proteins) are chosen for demonstration. Their corresponding average RMSD (aRMSD) values from the native structures are 0.93, 1.88, and 3.68 Å, respectively. Only the accuracy of χ_1 is used as an indicator. It is clear that OPUS-Rota2 and OPUS-Rota2i deliver better predicted values for χ_1 than all other methods.

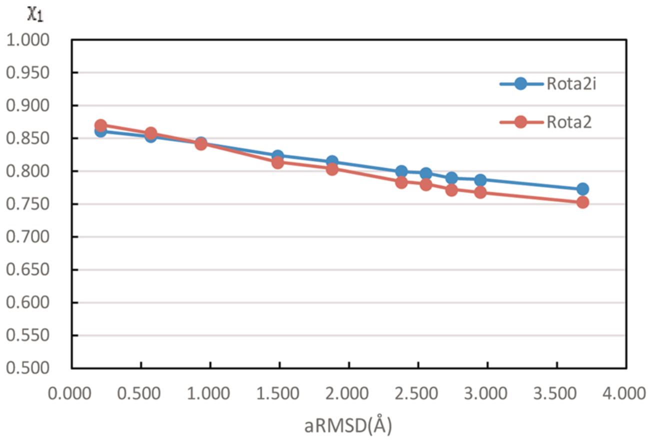


Figure 2. Performance of OPUS-Rota2 and OPUS-Rota2i on 10 different non-native test sets as a function of aRMSD. It suggests that the accuracy of prediction of χ_1 only decreases moderately when the main-chain aRMSD increases multiple times. Also, at larger values of aRMSD, OPUS-Rota2i seems to be better than OPUS-Rota2.

■ CONCLUDING DISCUSSION

Side-chain modeling is essential for protein structure prediction and high-accuracy refinement. However, despite the great success of the current methods, many of them are still relatively inefficient, which dramatically limits their usage. In this paper, we proposed a new side-chain modeling method called OPUS-Rota2. It has mainly two virtues: First, it balances the accuracy and the speed. Second, it also works well on non-native main chains. Both of these virtues are important in real applications, in which cases the main chains are usually non-native and the modeling process is iterative.

OPUS-Rota2 is based on our previous work, OPUS-Rota,[43] but with important modifications. The main difference is that we use the OPUS-DASF term, which specifically describes the relative conformation of side chains, to replace the OPUS-PSP[44] term in OPUS-Rota. In OPUS-Rota, the dominant contribution term for prediction accuracy is the vdW term, while in OPUS-Rota2, the dominant term becomes the OPUS-DASF term (Table 4). OPUS-DASF is based on our previous work, OPUS-CSF,[46] and it is a fast and accurate scoring function that can be used to distinguish the native protein structures from their decoys based on the conformations of their side chains exclusively. The good performance of OPUS-DASF (Table 2) is the critical methodological improvement in OPUS-Rota2 for modeling side chains.

We used both the 65-protein test set (DB65) in the OPUS-Rota paper[43] and the 379-protein test set (DB379) in the SCWRL4 paper[37] to evaluate the performance of OPUS-Rota2 and other methods. When the main chain is native, OPUS-Rota2 is more accurate than OPUS-Rota, SCWRL4, and OSCAR-star, but slightly less accurate than OSCAR-o (Table 3). However, OPUS-Rota2 is significantly faster than any other method: 2× faster than SCWRL4, 3× faster than OPUS-Rota, and 6× and 500× faster than OSCAR-star and OSCAR-o, respectively. Moreover, the nonsampling version OPUS-Rota2i is even 2× faster than OPUS-Rota2 while achieving a comparable accuracy with OSCAR-star.

To evaluate the performance of different side-chain modeling methods on the non-native main chains, we add Gaussian noise to the torsional angles of the original native main chains to build 10 non-native main-chain test sets. When the main chain is non-native, OPUS-Rota2i and OPUS-Rota2 are more accurate than any other method (Figure 1). It is noteworthy that OPUS-Rota2 is better than OPUS-Rota2i when the main chain is close to the native structure (<0.7 Å), and OPUS-Rota2i outperforms OPUS-Rota2 when the main chain is far from the native structure (>0.7 Å). This seems to suggest that, at an earlier stage of side-chain modeling, in which case the main chain is not close to the native state, OPUS-Rota2i is advantageous, while in a later stage when the main chain is more accurate, OPUS-Rota2 should be used. The main difference between OPUS-Rota2i and OPUS-Rota2 is that the former does not consider the vdW interactions between side-chain atoms on different residues, which is also the reason that OPUS-Rota2i is 2× faster than OPUS-Rota2.

Overall, OPUS-Rota2i and OPUS-Rota2 can construct the side-chain conformation accurately in a very short time, and the high speed in side-chain modeling plays a very important role as the modeling processes of protein structures are usually iterative and require many rounds of building and rebuilding. Thus, the combination of higher accuracy and speed of OPUS-Rota2i and OPUS-Rota2 in modeling side chains on both the native and non-native main chains makes them very useful tools in protein structure modeling.

In protein folding, the uniqueness of the native structure is solely determined by the side-chain packing specificity rather than by that of the main chain. Therefore, there are reasons to believe that the *nativeness* of side chains before the polypeptide chain reaches the final native state is vitally important to the final stage of folding; i.e., the ultimate folding process is driven by side-chain packing. However, most of protein structure prediction procedures, especially the knowledge-based homology modeling methods, build the main chains first and add the side chains last. Such a common practice is opposite to nature's law of folding. In other words, the ability of modeling the side-chain conformation based on not-so-native main-chains is extremely important in structure prediction, and it may potentially open up a totally new way of doing structure prediction.

Accessibility of OPUS-Rota2. The side-chain modeling program is freely available for all of the academic community.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jpma@bcm.edu.

ORCID

Jianpeng Ma: 0000-0003-2943-0779

Author Contributions

∇G.X. and T.M. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

J.M. acknowledges the support from grants by the National Institutes of Health (R01-GM127628, R01-GM116280) and the Welch Foundation (Q-1512). Q.W. wishes to acknowledge the support from grants by the National Institutes of Health (R01-GM127628, R01-GM116280), the Welch Foundation (Q-1826), and the Gillson-Longenbaugh Foundation. The work also partially benefited from Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZJLab.

REFERENCES

(1) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294* (5540), 93−96.

(2) Ponder, J. W.; Richards, F. M. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987**, *193* (4), 775−791.

(3) Dunbrack, R. L., Jr; Karplus, M. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.* **1993**, *230* (2), 543−574.

(4) Dunbrack, R. L., Jr; Cohen, F. E. Bayesian statistical analysis of protein side chain rotamer preferences. *Protein Sci.* **1997**, *6* (8), 1661−1681.

(5) De Maeyer, M.; Desmet, J.; Lasters, I. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des* **1997**, *2* (1), 53−66.

(6) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins: Struct., Funct., Genet.* **2000**, *40* (3), 389−408.

(7) Larriva, M.; Rey, A. Design of a rotamer library for coarse-grained models in protein-folding simulations. *J. Chem. Inf. Model.* **2014**, *54* (1), 302−313.

(8) Towse, C.-L.; Rysavy, S. J.; Vulovic, I. M.; Daggett, V. New dynamic rotamer libraries: data-driven analysis of side-chain conformational propensities. *Structure* **2016**, *24* (1), 187−199.

(9) Goldstein, R. F. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **1994**, *66* (5), 1335−1340.

(10) Gordon, D. B.; Mayo, S. L. Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure* **1999**, *7* (9), 1089−1098.

(11) Pierce, N. A.; Spriet, J. A.; Desmet, J.; Mayo, S. L. Conformational splitting: A more powerful criterion for dead end elimination. *J. Comput. Chem.* **2000**, *21* (11), 999−1009.

(12) Looger, L. L.; Hellinga, H. W. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics1. *J. Mol. Biol.* **2001**, *307* (1), 429−445.

(13) Desmet, J.; Spriet, J.; Lasters, I. Fast and accurate side chain topology and energy refinement (FASTER) as a new method for

protein structure optimization. *Proteins: Struct., Funct., Genet.* **2002**, *48* (1), 31−43.

(14) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L., Jr A graph theory algorithm for rapid protein side chain prediction. *Protein Sci.* **2003**, *12* (9), 2001−2014.

(15) Chazelle, B.; Kingsford, C.; Singh, M. A semidefinite programming approach to side chain positioning with new rounding strategies. *INFORMS Journal on Computing* **2004**, *16* (4), 380−392.

(16) Kingsford, C. L.; Chazelle, B.; Singh, M. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* **2005**, *21* (7), 1028−1039.

(17) Xu, J. Rapid protein side-chain packing via tree decomposition. In *Research in Computational Molecular Biology*; Annual International Conference on Research in Computational Molecular Biology; Springer, 2005; pp 423−439.

(18) Xie, W.; Sahinidis, N. V. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **2006**, *22* (2), 188−194.

(19) Chong, K. F.; Leong, H. W. A Merge-Decoupling Dead End Elimination algorithm for protein side-chain conformation. *International journal of data mining and bioinformatics* **2007**, *1* (4), 372−388.

(20) Georgiev, I.; Donald, B. R. Dead-end elimination with backbone flexibility. *Bioinformatics* **2007**, *23* (13), No. i185.

(21) Hartmann, C.; Antes, I.; Lengauer, T. IRECS: A new algorithm for the selection of most probable ensembles of side chain conformations in protein models. *Protein Sci.* **2007**, *16* (7), 1294−1307.

(22) Santana, R.; Larrañaga, P.; Lozano, J. A. Side chain placement using estimation of distribution algorithms. *Artif. Intell. Med.* **2007**, *39* (1), 49−63.

(23) Spassov, V. Z.; Yan, L.; Flook, P. K. The dominant role of side chain backbone interactions in structural realization of amino acid code. ChiRotor: A side chain prediction algorithm based on side chain backbone interactions. *Protein Sci.* **2007**, *16* (3), 494−506.

(24) Burley, K. H.; Gill, S. C.; Lim, N. M.; Mobley, D. L. Enhancing Sidechain Rotamer Sampling Using Non-Equilibrium Candidate Monte Carlo. *J. Chem. Theory Comput.* **2019**, *15*, 1848.

(25) Xiang, Z.; Honig, B. Extending the accuracy limits of prediction for side-chain conformations1. *J. Mol. Biol.* **2001**, *311* (2), 421−430.

(26) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320* (3), 597−608.

(27) Liang, S.; Grishin, N. V. Side chain modeling with an optimized scoring function. *Protein Sci.* **2002**, *11* (2), 322−331.

(28) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **2003**, *331* (1), 281−299.

(29) Eyal, E.; Najmanovich, R.; Mcconkey, B. J.; Edelman, M.; Sobolev, V. Importance of solvent accessibility and contact surfaces in modeling side chain conformations in proteins. *J. Comput. Chem.* **2004**, *25* (5), 712−724.

(30) Peterson, R. W.; Dutton, P. L.; Wand, A. J. Improved side chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci.* **2004**, *13* (3), 735−751.

(31) Leaver Fay, A.; Butterfoss, G. L.; Snoeyink, J.; Kuhlman, B. Maintaining solvent accessible surface area under rotamer substitution for protein design. *J. Comput. Chem.* **2007**, *28* (8), 1336−1341.

(32) Lopes, A.; Alexandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins: Struct., Funct., Genet.* **2007**, *67* (4), 853−867.

(33) Xiang, Z.; Steinbach, P. J.; Jacobson, M. P.; Friesner, R. A.; Honig, B. Prediction of side chain conformations on protein surfaces. *Proteins: Struct., Funct., Genet.* **2007**, *66* (4), 814−823.

(34) Zhu, K.; Shirts, M. R.; Friesner, R. A. Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects. *J. Chem. Theory Comput.* **2007**, *3* (6), 2108−2119.

(35) Liang, S.; Zheng, D.; Zhang, C.; Standley, D. M. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **2011**, *27* (20), 2913−2914.

(36) Liang, S.; Zhou, Y.; Grishin, N.; Standley, D. M. Protein side chain modeling with orientation dependent atomic force fields derived by series expansions. *J. Comput. Chem.* **2011**, *32* (8), 1680−1686.

(37) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. Improved prediction of protein side chain conformations with SCWRL4. *Proteins: Struct., Funct., Genet.* **2009**, *77* (4), 778−795.

(38) Mendes, J.; Baptista, A. M.; Carrondo, M. A.; Soares, C. M. Improved modeling of side chains in proteins with rotamer based methods: A flexible rotamer model. *Proteins: Struct., Funct., Genet.* **1999**, *37* (4), 530−543.

(39) Wang, C.; Schueler Furman, O.; Baker, D. Improved side chain modeling for protein-protein docking. *Protein Sci.* **2005**, *14* (5), 1328−1339.

(40) Jain, T.; Cerutti, D. S.; McCammon, J. A. Configurational bias sampling technique for predicting side chain conformations in proteins. *Protein Sci.* **2006**, *15* (9), 2029−2039.

(41) Zhang, W.; Duan, Y. Grow to Fit Molecular Dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement. *Protein Eng., Des. Sel.* **2006**, *19* (2), 55−65.

(42) Cao, Y.; Song, L.; Miao, Z.; Hu, Y.; Tian, L.; Jiang, T. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. *Bioinformatics* **2011**, *27* (6), 785−790.

(43) Lu, M.; Dousis, A. D.; Ma, J. OPUS Rota: A fast and accurate method for side chain modeling. *Protein Sci.* **2008**, *17* (9), 1576−1585.

(44) Lu, M.; Dousis, A. D.; Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* **2008**, *376* (1), 288−301.

(45) Xu, G.; Ma, T.; Zang, T.; Sun, W.; Wang, Q.; Ma, J. OPUS-DOSP: A Distance-and Orientation-Dependent All-Atom Potential Derived from Side-Chain Packing. *J. Mol. Biol.* **2017**, *429* (20), 3113−3120.

(46) Xu, G.; Ma, T.; Zang, T.; Wang, Q.; Ma, J. OPUS CSF: A C atom based scoring function for ranking protein structural models. *Protein Sci.* **2018**, *27* (1), 286−292.

(47) Shapovalov, M. V.; Dunbrack, R. L., Jr A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19* (6), 844−858.

(48) Tien, M. Z.; Sydykova, D. K.; Meyer, A. G.; Wilke, C. O. PeptideBuilder: A simple Python library to generate model peptides. *PeerJ* **2013**, *1*, No. e80.

(49) Walker, M. W.; Shao, L.; Volz, R. A. Estimating 3-D location parameters using dual number quaternions. *CVGIP: image understanding* **1991**, *54* (3), 358−367.

(50) Zhou, H.; Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **2011**, *101* (8), 2043−2052.

(51) Xu, G.; Ma, T.; Wang, Q.; Ma, J. OPUS SSF: A Side chain inclusive Scoring Function for Ranking Protein Structural Models. *Protein Sci.* **2019**, *28* (6), 1157−1162.

(52) Deng, H.; Jia, Y.; Zhang, Y. 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **2016**, *32* (3), 378−387.

(53) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct., Funct., Genet.* **2003**, *53* (1), 76−87.

(54) Zhang, J.; Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **2010**, *5* (10), No. e15386.