BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 22 2011, pages 3117–3122
doi:10.1093/bioinformatics/btr538

Structural bioinformatics

Advance Access publication September 23, 2011

RASP: rapid modeling of protein side chain conformations

Zhichao Miao[1,2], Yang Cao[1,2] and Taijiao Jiang[1,*]

[1]National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101 and [2]Graduate University of Chinese Academy of Sciences, Beijing 100039, China

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Modeling of side chain conformations constitutes an indispensable effort in protein structure modeling, protein–protein docking and protein design. Thanks to an intensive attention to this field, many of the existing programs can achieve reasonably good and comparable prediction accuracy. Moreover, in our previous work on CIS-RR, we argued that the prediction with few atomic clashes can complement the current existing methods for subsequent analysis and refinement of protein structures. However, these recent efforts to enhance the quality of predicted side chains have been accompanied by a significant increase of computational cost.

Results: In this study, by mainly focusing on improving the speed of side chain conformation prediction, we present a RApid Side-chain Predictor, called RASP. To achieve a much faster speed with a comparable accuracy to the best existing methods, we not only employ the clash elimination strategy of CIS-RR, but also carefully optimize energy terms and integrate different search algorithms. In comprehensive benchmark testings, RASP is over one order of magnitude faster (∼40 times over CIS-RR) than the recently developed methods, while achieving comparable or even better accuracy.

Availability: RASP is available to non-commercial users at our website: http://jianglab.ibp.ac.cn/lims/rasp/rasp

Contact: taijiao@moon.ibp.ac.cn

Supplementary Information: Supplementary information is available at *Bioinformatics* online.

Received on July 12, 2011; revised on August 24, 2011; accepted on September 4, 2011

1 INTRODUCTION

Protein side chain conformation prediction or side chain packing is a crucial step in both protein design (Dahiyat and Mayo, 1996; Fromer *et al.*, 2010; Jones, 1994) and protein structure modeling (Bower *et al.*, 1997; Holm and Sander, 1991; Rohl *et al.*, 2004). As such, during the last two decades, many efforts have been dedicated to the prediction of protein side chain conformations, leading to the development of many side chain packing programs (Canutescu *et al.*, 2003; DeMaeyer *et al.*, 1997; Dunbrack and Karplus, 1993; Fromer *et al.*, 2010; Hartmann *et al.*, 2007; Jiang *et al.*, 2011; Krivov *et al.*, 2009; Liang and Grishin, 2002; Liang *et al.*, 2011; Lu *et al.*, 2008; Mcgregor *et al.*, 1987; Ponder and Richards, 1987; Tuffery *et al.*, 1991; Xiang and Honig, 2001; Xu, 2005). In general, these side chain packing programs formulate the side chain conformation prediction problem as a combinatorial search problem based on discrete side chain conformations, called rotamers. The problem usually involves three key elements, summarized as follows:

(i) A rotamer library of discrete side chain conformations (Dunbrack, 2002; Dunbrack and Cohen, 1997; Shapovalov and Dunbrack, 2011). Although some approaches attempted to model the continuous side chain conformations (Feyfant *et al.*, 2007; Harder *et al.*, 2010), most of the current methods rely on discrete rotamers, which can significantly reduce computational expense (Peterson *et al.*, 2004).

(ii) An energy function for rotamer selection. A number of energy functions for side chain packing have been proposed, ranging from simple van de Waals potential (Bower *et al.*, 1997; DeMaeyer *et al.*, 1997; Vasquez, 1995) to more complicated potentials by incorporating hydrogen bonding term (Krivov *et al.*, 2009), solvation term (Jacobson *et al.*, 2002; Mendes *et al.*, 2001) and statistical orientation term (Liang *et al.*, 2011; Lu *et al.*, 2008) to improve the prediction accuracy.

(iii) A search algorithm. The use of rotamer library leads to the formulation of the side chain packing problem as a combinatorial problem, which finds the best solution from all the possible combinations constrained by side chain rotamers. Many search algorithms have been proposed to solve the combinatorial problem, including dead-end elimination (DEE) (Desmet *et al.*, 1992), simulated annealing (Lee and Subbiah, 1991), Monte Carlo (Gray *et al.*, 2003), A* (Leach and Lemon, 1998), integer programming (Kingsford *et al.*, 2005), self-consistent mean field (Lee, 1994; Mendes *et al.*, 1999) and graph theory-based approach (Canutescu *et al.*, 2003; Samudrala and Moult, 1998). The combining of these search algorithms is critical in side chain prediction. For example, to achieve a fast speed, SCWRL4 (Krivov *et al.*, 2009) and SCATD (Xu, 2005) combine DEE, branch-and-bound and tree-decomposition search.

As summarized above, the appropriate consideration of the above three elements is critical in developing rotamer-based side chain packing programs. Thanks to the intensive previous efforts, the prediction of side chain conformations has become more and more accurate. However, the improvement of accuracy usually comes with the increase of computational time. For example, although the recently developed side chain packing program SCRWL4 from Dunbrack's laboratory and the program CIS-RR developed in our laboratory (Jiang *et al.*, 2011) have an improvement of ∼3% in χ accuracy over SCRWL3, their speed is over 6 times slower than SCRWL3, indicating the challenge in achieving both high accuracy and high speed in the prediction of protein side chain conformations. In this study, by carefully considering the key elements summarized

*To whom correspondence should be addressed.

above, we present a *RA*pid Side-chain Predictor, called RASP, which builds the protein side chains over one order of magnitude faster than the best existing ones while achieving comparable accuracy.

2 METHODS

Similar to CIS-RR, RASP uses clash detection-guided side chain optimization to alleviate atomic clashes caused by rigid rotamer approximation. But unlike CIS-RR which couples the elimination of atomic clashes with the process of side chain packing in an iterative search, RASP efficiently eliminates atomic clashes after the generation of a high-quality structure. Therefore, in RASP, we focus on (i) rapid generation of high-quality initial structures by carefully considering the key elements of rotamer-based side chain packing algorithms, which is described in Section 2.1 and (ii) rapid elimination of atomic clashes by relaxing those residues in clashes, which is described in Section 2.2.

2.1 Rapid prediction of side chain conformations with high accuracy

2.1.1 Rotamer building Rotamer dihedrals along with their probabilities are read from a binary-formatted backbone-dependent rotamer library (Dunbrack and Cohen, 1997) in order from highest to lowest probability until the cumulative probability reaches 98%. Then, the coordinates of all side chain atoms are built according to rotamer dihedrals (Parsons *et al.*, 2005) using standard side chain topology (Engh and Huber, 1991).

2.1.2 Energy calculation Assuming that the rotamer r_i is one of the all possible rotamers of residue i, the total energy of a protein system of N residues is expressed as:

$$E_{\text{total}} = \sum_{i}^{N} E_{\text{lib}}(r_i) + \sum_{i}^{N}\sum_{j} E_{bb-sc}(r_i, j) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E_{sc-sc}(r_i, r_j) \quad (1)$$

The side chain packing problem is a combinatorial problem, which finds the set of r_i $(i=1,\ldots,N)$ that gives the lowest E_{total} over all possible rotamers of a residue.

Rotamer probability term, $E_{\text{lib}}(r_i)$, takes the same form as the one used in SCWRL4 (Krivov *et al.*, 2009):

$$E_{\text{lib}}(r_i) = -w_{aa} \log \frac{p(r_i, \phi\psi)}{p(r_{\max}, \phi\psi)} \quad (2)$$

Given backbone dihedrals (Φ and Ψ), this term expresses the relative probability of a rotamer r_i to the highest probability rotamer $r_{\max}$. Scaling factor w_{aa} is residue-type dependent (Supplementary Table S1).

Besides, three other energy terms are used to characterize the interactions between atoms (both E_{bb-sc} and E_{sc-sc}): van de Waals potential, disulfide term and hydrogen bonding term:

$$\mathrm{E}_{bb-sc/sc-sc} = E_{vdW} + \mathrm{E}_{SS} + \mathrm{E}_{[O,H]} \quad (3)$$

The van de Waals potential, E_{vdw}, is adapted from the one used in OPUS-PSP (Jain *et al.*, 2006; Lu *et al.*, 2008), which is formulated as follows:

$$E_{vdW}(i,j) = \begin{cases} 50e_{ij} & \text{if } d' < 0.465 \\ e_{ij}(80 - 64.5d') & \text{if } 0.465 \le d' < 0.75 \\ 1.63e_{ij}\left[\left(\frac{1}{d'}\right)^{12} - 2\left(\frac{1}{d'}\right)^{6}\right] & \text{if } 0.75 \le d' < 0.8929 \\ 0.99e_{ij}\left[\left(\frac{1}{d'}\right)^{12} - 2\left(\frac{1}{d'}\right)^{6}\right] & \text{if } 0.8929 \le d' < 2.3 \end{cases} \quad (4)$$

where $e_{ij} = \sqrt{e_i e_j}$, and e_i, e_j are well-depths from charmm19 (Brooks *et al.*, 1983). $d' = d_{ij}/R_{ij}$, d_{ij} is distance between atoms i and j. R_{ij} is summation of atomic radii (Supplementary Table S2) of the two atoms i and j. 1.63 and 0.99 are scaling factors to express the difference between repulsive and attractive effect. For the repulsive term, it is capped at a maximum value of $50e_{ij}$ to alleviate fixed rotamer approximation.

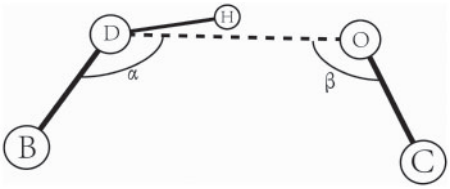

Fig. 1. Hydrogen bonds between hydroxyl and carboxyl. B is the Base of the hydrogen donor (base of hydroxyl) and D is the hydrogen donor (O atom in hydroxyl). O is hydrogen acceptor (O atom in carboxyl) and C is its base C atom in carboxyl. α is the angle between hydrogen acceptor, hydrogen donor and base of the hydrogen donor; β is the angle between hydrogen donor, hydrogen acceptor and base of the hydrogen acceptor.

The disulfide term, E_{SS}, is a simplified version of the one used in SCWRL3 (Canutescu *et al.*, 2003):

$$E_{SS} = 6\left(\left|d - 2.06\text{Å}\right| + \frac{|A_1 - 105^\circ| + |A_2 - 105^\circ|}{100} + \frac{||\chi_3| - 90^\circ|}{140}\right) - 11.4 \quad (5)$$

where d is Sγ–Sγ distance, A_1 and A_2 are two Sγ–Sγ–Cβ angles, χ_3 is the Cβ–Sγ–Sγ–Cβ dihedral angle. The 6, 100 and 140 are scaling factors. The bond energy of standard disulfide takes 11.4 kcal/mol.

The hydrogen bonding term, $E_{[O,H]}$, only considers hydrogen bonds between hydroxyl and carboxyl, which is formulated as follows:

$$E_{[O,H]} = -1.8\sqrt{\frac{(\cos(\alpha - 111.5) - \cos 37)(\cos(\beta - 120) - \cos 47)}{(1 - \cos 37)(1 - \cos 47)}} \quad (6)$$

The description of the term is illustrated in Figure 1. This term is calculated only if the distance between an O atom in the carboxyl and the O atom in the hydroxyl is <3.2 Å. Since no explicit hydrogen atom coordinates are needed in this term, its calculation is relatively very fast.

In order to efficiently compute the interacting energies between residues within a protein system, we only consider the pairs of residues that have effective contact. A pair of residues are assumed to have effective contact if the Cβ atom of one side chain falls within a region of a hemi-sphere centered at the Cβ atom of another side chain (illustration of side chain in effective contact see Fig. 2):

$$\begin{aligned} \text{Pair}(i,j) = \text{contact, if } & \mathrm{d}(C\beta_i, C\beta_j) < r_i + r_j + 5\text{Å} \\ \text{and } & (\angle C\alpha_i C\beta_i C\beta_j > 90^\circ \text{ or } \angle C\alpha_j C\beta_j C\beta_i > 90^\circ) \end{aligned} \quad (7)$$

where $d(C\beta_i, C\beta_j)$ is the distance between the two Cβ atoms, r_i, r_j are radii of side chain hemisphere of residue i and j, respectively.

To prevent a side chain from severely colliding with backbone atoms, one rotamer is excluded when its backbone energy [E_{bb-sc}, second term in Equation (1)] is 8.0 U higher than any other rotamers of the same residue.

2.1.3 Combinatorial search To achieve a fast and convergent search, we employed a combination of DEE algorithm (Desmet *et al.*, 1992), graph theory-based search (Samudrala and Moult, 1998), branch-and-terminate search (Gordon and Mayo, 1999), Monte Carlo search (Liu, 2008) and backtrack algorithm (Tarjan, 1972). The DEE and graph-theory based search were used by following a similar procedure used in SCWRL3 (Canutescu *et al.*, 2003): a simple Goldstein DEE algorithm (Goldstein, 1994) was first used to reduce the combinatorial space, and then an interaction graph was constructed and further divided into bi-connected components that can be solved by branch-and-terminate search strategy. To speed up the search process, we made two modifications described as follows:

(i) In our graph theory-based search, the graph was constructed by connecting the residue pairs whose energy difference between any

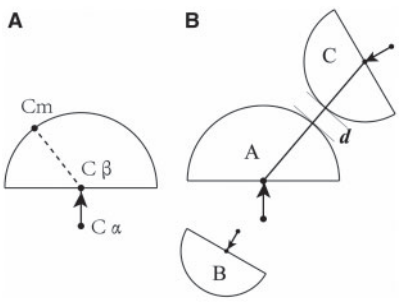

Fig. 2. Definition of effective contact between residues. (**A**) The side chain of a residue is represented as a hemisphere, whose center is $C\beta$ atom and $C\alpha$-$C\beta$ is the direction. Cm is the most remote atom on the side chain. The radii of different side chain hemispheres (Supplementary Table S3). As seen from Supplementary Table S3, a residue with longer side chain has longer radius of hemisphere, derived from the training set (the longest distance from a side chain atom to the $C\beta$ atom of the side chain is taken as radius of the hemisphere). (**B**) Case of residue pairs in effective contact or not in effective contact. Residue A and C form a pair in effective contact when ***d*** is shorter than 5 Å, A and B do not have effective contact.

rotamer combinations of the residue pair is bigger than a threshold of 3 kcal/mol:

$$\text{Edge}(i,j) = \text{contact, if } \max(E_{sc-sc}(i,j)) - \min(E_{sc-sc}(i,j)) > 3\,\text{kcal/mol} \quad (8)$$

This is because of the introduction of attractive terms in RASP. In SCWRL3, only the rotamer pairs that have repulsive effects are considered to form 'edges' in the graph. While in RASP, we considered edges for the rotamer pairs having either repulsive effects or attractive effects. However, the introduction of attractive term led to more 'edges' (residue pairs in contact) in the graph, and making the graph more complicated and difficult to solve.

(ii) Therefore, to effectively solve the bi-connected components, we used two strategies. For the bi-connected components with a combination number over 10^{15}, a simulated annealing Monte Carlo is carried out. Otherwise, the branch-and-terminate search strategy is performed (Gordon and Mayo, 1999).

The MC search starts from a structure with all residues using the rotamer of the lowest self energy. The acceptance probability of a new rotamer n to replace an old rotamer o, denoted as $p(o \to n)$, is calculated as follows:

$$p(o \to n) = \exp(-\Delta E_{\text{tot}}(o \to n)/T) \quad (9)$$

First, rotamers with low acceptance probability ($p(o \to n) < \exp(-10)$) are eliminated. Then, it performs 100 rounds of standard Monte Carlo search with a gradual temperature decrease from 2 to 0.02 U followed by three rounds of greedy search.

2.2 Clash detection-guided rotamer relaxation

As defined in CIS-RR, two atoms are deemed to collide when the distance between them is <60% the sum of their van der Waals radii. Residues with no clashes are kept intact, others are relaxed using the CIS-RR (Jiang *et al.*, 2011) approach.

2.3 Training and evaluation

2.3.1 Training and parameterization The training and testing dataset were obtained from PISCES server (Wang and Dunbrack, 2003) with resolution ≤1.8 Å, R-factor <25% and mutual sequence identity <25% (Supplementary Table S4). After elimination of the structures with incomplete side chain, 300 structures not present in SCWRL4 test set described below were used as training data, and the other 2412 structures as test data.

By maximizing the summation of per cent correct for χ_1 and χ_{1+2} on the training data of 300 structures, we optimized the parameters of individual terms [Equations (2)–(6)] including scaling factors for each term, van de Waals radii in van de Waals potential (Supplementary Table S2), parameters in disulfide term, hydrogen bonding term and MC search times. The parameters of van de Waals potential [Equation (4)] were first optimized over the training set. Then the other terms were optimized one by one using a greedy algorithm.

2.3.2 Evaluation Two different test sets are used for evaluation of RASP and other programs. One test set consists of 2412 protein structures described above (see Section 2.3.1 and Supplementary Table S5). This is a very comprehensive test set, which contains 437 393 residues, namely at least 6700 counts per residue type. The other is taken from SCWRL4 (Krivov *et al.*, 2009), thus called SCRWRL4 test set, which consists of 379 structures (58 231 residues in total).

RASP and some other programs are evaluated on the above two test sets using their respective default settings. We use two criteria to assess the side chain packing accuracy. One is percent correct of χ_1 and χ_{1+2}. If the dihedrals χ_1 and χ_2 of a modeled side chain are within 40° those of the native side chain conformation, they can be regarded to be correct (Bower *et al.*, 1997). The other is side chain atom root mean square deviation (RMSD), which is computed as follows:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} d^2}{N}} \quad (10)$$

d is the distance between a native coordinate and the predicted one. N is the total atom number.

As defined in CIS-RR (Jiang *et al.*, 2011), two atoms are assumed to be in clash if the distance between them is <60% of the sum of their van de Waals radii taken from Rosetta program (Rohl *et al.*, 2004), and the performance of elimination of atomic clashes is quantified based on the number of atom pairs in clash.

2.4 Implementation and software availability

RASP was implemented in object-oriented C++, compiled using gcc version 4 with -O3 option. It has been tested on five versions of linux platforms. For fast computation, the pair-wise energies are stored in a 4D array for quick indexing. The evaluation of the program was carried out on Intel Q9550 processor. The binary executable program is freely available to non-profit research via http://jianglab.ibp.ac.cn/lims/rasp/rasp. All datasets are available at the same web. Commercial users should contact the investigator for consent.

3 RESULTS

3.1 RASP has a comparable prediction accuracy but is much faster than CIS-RR

To evaluate the performance of RASP by comparing to CIS-RR, the two programs were tested on a comprehensive test set consisting of 2412 high-resolution X-ray structures (see Section 2.3). The test showed that both programs have very close performance in prediction accuracy: 86.02% for RASP versus 85.49% for CIS-RR for the percent correct of χ_1 and 75.92% for RASP versus 75.68% for CIS-RR for the percent correct of χ_{1+2}. Both programs also have similar prediction performance for different residue types (Supplementary Fig. S1, Tables S6 and S7). Moreover, both programs can effectively eliminate atomic clashes. There exists one pair of atoms in clash in about 4 (2412 proteins/642 clashes) protein structures modeled by CIS-RR and only in about 5 (2412 proteins/495 clashes) protein structures modeled by RASP. However, RASP is much faster (~40 times) than CIS-RR.

Table 1. Comparison of RASP with some recently developed side chain prediction programs on SCWRL4 test set

Program	Time	Clash	χ_1 (%)	χ_{1+2} (%)	RMSD (Å)
RASP	1 min 47 s	47	85.10	74.71	1.47
CIS-RR	73 min 55 s	59	84.88	74.88	1.47
SCWRL4	33 min 24 s	411	85.03	75.44	1.46
SCWRL3	5 min 8 s	1107	82.17	71.26	1.58
OPUS-Rota	26 min 33 s	623	85.03	75.05	1.43
IRECS	38 min 25 s	1201	83.56	71.74	1.66

Percent correct of χ_1 is defined as the percentage of residues whose predicted χ_1 dihedral is within 40° of the χ_1 dihedral of native side chains, while Percent χ_{1+2} correct is defined as the percentage of residues for which both χ_1 and χ_2 are within 40° of those of native side chains.

RASP finished the prediction of all 2412 protein structures within 16 min of CPU time, namely <0.4 s per protein£¬While CIS-RR spent 601 min on the 2412 protein structures.

3.2 Comparison of RASP with other popular side chain modeling programs

We further compared RASP with some recently developed programs, including CIS-RR (Jiang *et al.*, 2011), SCWRL4 (Krivov *et al.*, 2009), OPUS-Rota (Lu *et al.*, 2008) and IRECS (Hartmann *et al.*, 2007), on the SCWRL4 test set. As shown in Table 1, for prediction accuracy in terms of percent correct for χ_1 and χ_{1+2} and RMSD, RASP is comparable to the recently developed side chain programs (detailed data can be found in Supplementary Table S7). While for speed, RASP is relatively much faster, being 14 times faster than OPUS-Rota, 18 times faster than SCWRL4 and 40 times faster than CIS-RR. Moreover, it generates fewer clashes than CIS-RR, SCWRL4 and OPUS-Rota do.

We further investigated the distribution of prediction time by RASP, CIS-RR, SCWRL3 and SCWRL4 on the SCWRL4 test set. Figure 3 shows that RASP completed the predictions within 1 s for nearly all proteins (370 out of the 379 proteins in the SCWRL4 test set), showing significant advantage in speed over the other three programs.

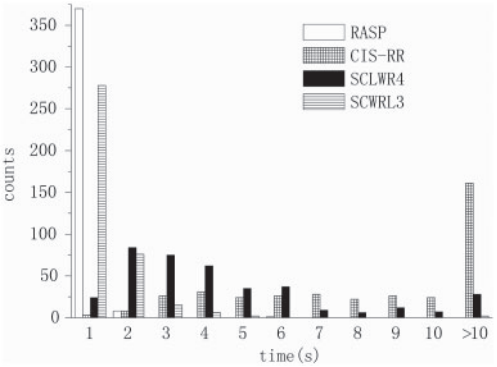

Fig. 3. Comparison of the prediction time distribution for RASP, CIS-RR, SCWRL4 and SCRWRL3. The *X*-axis is the prediction time by seconds (s), and *Y*-axis is the number of structures (counts) finished within a given time.

4 DISCUSSION

Determination of side chain conformations on a fixed protein backbone plays an important role in protein structure prediction, protein design and molecular docking. In hitherto, many methods have been developed to predict protein side chain conformations (Canutescu *et al.*, 2003; DeMaeyer *et al.*, 1997; Dunbrack and Karplus, 1993; Fromer *et al.*, 2010; Hartmann *et al.*, 2007; Jiang *et al.*, 2011; Krivov *et al.*, 2009; Liang and Grishin, 2002; Liang *et al.*, 2011; Lu *et al.*, 2008; Mcgregor *et al.*, 1987; Ponder and Richards, 1987; Tuffery *et al.*, 1991; Xiang and Honig, 2001; Xu, 2005). Although the prediction has become more and more accurate, the gradual improvement of accuracy usually comes with the dramatic increase of computational cost. In this study, in order to improve speed without at expense of prediction accuracy, we have developed a more powerful program, called RASP. The tests showed that RASP achieves high prediction accuracy comparable to the best existing methods, but is much faster.

The good performance of RASP lies in its elegant integration of the strategies used in the existing approaches, which is contributed by three critical points discussed as follows:

One is the design of energy function to achieve both high accuracy and high speed. Previous studies have indicated that van der Waals potential alone is able to achieve a high accuracy in prediction of side chain conformations (Vasquez, 1995), suggesting the dominant role it plays in side chain packing. Although the incorporation of other energy terms could improve the accuracy of side chain packing, the consideration of complicated energy terms would aggravate the computational expenses. Therefore, in developing an effective energy function for side chain packing, all the individual energy terms were carefully implemented for fast speed and high accuracy. To calculate van der Waals potential, the hydrogen atoms were not represented explicitly and their effects, through parameterization, could be captured in the heavy atoms that are linked to hydrogen atoms (Supplementary Table S2). As shown in Supplementary Table S2, among the heavy atoms of the same type, the ones with more hydrogen links have longer van de Waals radii. For simplicity and fast computation, the effect of dihedral $C\alpha$–$C\beta$–Sγ–Sγ is omitted in the disulfide term, and the hydrogen bonding term only depends on the orientation between the hydrogen donor and hydrogen acceptor. The careful consideration of these energy terms led us to develop an effective energy function for side chain packing. As demonstrated by our testing described above, RASP is not only much faster than CIS-RR, but also is slightly better than CIS-RR (with ~0.2% improvement in both χ_1 and χ_{1+2} accuracy) (Supplementary Table S6 and S7).

The second is the use of clash detection-guided rotamer relaxation after side chain packing process. Further elimination of the atomic clashes in protein structures with modeled side chains can lead to more harmonic structures. Our recently developed CIS-RR coupled side chain packing process with the elimination of atomic clashes by using clash detection-guided iterative search in rotamer relaxation. Although in CIS-RR both the accuracy of side chain packing and removal of atomic clashes can be achieved, it is time consuming. In RASP, we uncoupled the side chain packing process and atomic clash elimination process. The rotamer relaxation was performed on the side chains that are involved in atomic clashes. By doing this, we found that rotamer relaxation process in RASP can effectively

Table 2. The effect of clash detection-guided rotamer relaxation (RR) on the performance of RASP

RASP	Time	Clash	χ_1 (%)	χ_{1+2} (%)	RMSD (Å)
Without RR	1 min 18 s	677	85.07	74.67	1.48
With RR	1 min 47 s	47	85.10	74.71	1.47

eliminate the atomic clashes without incurring the computation time significantly (Table 2, Supplementary Table S8). Moreover, the elimination of atomic clashes has led to a slight increase of prediction accuracy.

The SCWRL4 test set was used for the evaluation. Total time for the prediction is evaluated. Clashes are those residue pairs whose distance is <60% the summation of their van de Waals radii. χ dihedrals within 40° are considered correct.

The third is the implementation of effective search algorithms. To achieve a fast and convergent search, we designed a combinatorial search strategy involving graph theory-based approach which decomposed the graph of contact residue pairs to bi-connected components (see Section 2). For a bi-connected component of N residues, its time complexity is $\sim O(m^N)$ (m is the average rotamer number in a residue) for an exhaustive search like branch-and-terminate approach, while for Monte Carlo search the time complexity is $O(m \times N)$.

Therefore, in design of search algorithms, we used MC search instead of the exhaustive branch-and-terminate search for a large bi-connected component with over 20 residues. Indeed, we found use of MC search can significantly reduce the computation time for large proteins. Supplementary Table S9 compares the performance of RASP using MC or not on SCWRL4 test set. As can be seen from Supplementary Table S9, although the improvement by adding MC is general to nearly all proteins, it varies significantly for different proteins. Especially, for those proteins that tend to form highly connected graphs (often occurring in large proteins of over 300 residues), MC contributes significantly to the speed improvement.

Taken together, RASP not only combines the advantages of the existing programs in both prediction accuracy and clash elimination, but also achieves a much faster speed. We believe RASP would be a very useful tool for fast side chain modeling that can complement the current existing methods in a wide range of applications.

ACKNOWLEDGEMENTS

We gratefully acknowledge Dr Roland L. Dunbrack for his help on detailed information of rotamer library and Drs Jianpeng Ma and Mingyang Lu for the benchmark dataset. We would like to thank Mrs Cuxia Chen for the help in building the website.

Funding: Bai Ren Project of Chinese Academy of Sciences and grants from the Chinese Ministry of Science and Technology; the National Science and Technology Key Project (2008ZX10004-013) and Project973(2009CB825506) (to T.J.).

Conflict of Interest: none declared.

REFERENCES

Bower,M.J. *et al.* (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.*, **267**, 1268–1282.

Brooks,B.R. *et al.* (1983) Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.

Canutescu,A.A. *et al.* (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.

Dahiyat,B.I. and Mayo,S.L. (1996) Protein design automation. *Protein Sci.*, **5**, 895–903.

DeMaeyer,M. *et al.* (1997) All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des.*, **2**, 53–66.

Desmet,J. *et al.* (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**, 539–542.

Dunbrack,R.L. (2002) Rotamer libraries in the 21(st) century. *Curr. Opin. Struct. Biol.*, **12**, 431–440.

Dunbrack,R.L. and Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, **6**, 1661–1681.

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

Engh,R.A. and Huber,R. (1991) Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement. *Acta Crystallogr. A*, **47**, 392–400.

Feyfant,E. *et al.* (2007) Modeling mutations in protein structures. *Protein Sci.*, **16**, 2030–2041.

Fromer,M. *et al.* (2010) SPRINT: side-chain prediction inference toolbox for multistate protein design. *Bioinformatics*, **26**, 2466–2467.

Goldstein,R.F. (1994) Efficient Rotamer Elimination Applied to Protein Side-Chains and Related Spin-Glasses. *Biophys. J.*, **66**, 1335–1340.

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

Gray,J.J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Harder,T. *et al.* (2010) Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, **11**, -.

Hartmann,C. *et al.* (2007) IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.*, **16**, 1294–1307.

Holm,L. and Sander,C. (1991) Database Algorithm for Generating Protein Backbone and Side-Chain Coordinates from a C-Alpha Trace Application to Model-Building and Detection of Coordinate Errors. *J. Mol. Biol.*, **218**, 183–194.

Jacobson,M.P. *et al.* (2002) On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.*, **320**, 597–608.

Jain,T. *et al.* (2006) Configurational-bias sampling technique for predicting side-chain conformations in proteins. *Protein Sci.*, **15**, 2029–2039.

Jones,D.T. (1994) De-novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.*, **3**, 567–574.

Jiang,T.J. *et al.* (2011) Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. *Bioinformatics*, **27**, 785–790.

Kingsford,C.L. *et al.* (2005) Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, **21**, 1028–1036.

Krivov,G.G. *et al.* (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.

Leach,A.R. and Lemon,A.P. (1998) Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins Struct. Funct. Genet.*, **33**, 227–239.

Lee,C. (1994) Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.*, **236**, 918–939.

Lee,C. and Subbiah,S. (1991) Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, **217**, 373–388.

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

Liang,S.D. *et al.* (2011) Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. *J. Comput. Chem.*, **32**, 1680–1686.

Liu,J. (2008) *Monte Carlo Strategies in Scientific Computing*. 1st edn. Springer, 2001, 2nd printing, 2001, XVI, p 346. ISBN 978-0-387-76369-9.

Lu,M.Y. *et al.* (2008) OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.*, **376**, 288–301.

Lu,M.Y. *et al.* (2008) OPUS-Rota: A fast and accurate method for side-chain modeling. *Protein Sci.*, **17**, 1576–1585.

Mcgregor,M.J. *et al.* (1987) Analysis of the relationship between side-chain conformation and secondary structure in globular-proteins. *J. Mol. Biol.*, **198**, 295–310.

Mendes,J. *et al.* (1999) Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins Struct. Funct. Genet.*, **37**, 530–543.

Mendes,J. *et al.* (2001) Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. *J. Comput. Aid. Mol. Des.*, **15**, 721–740.

Parsons,J. *et al.* (2005) Practical conversion from torsion space to Cartesian space for in silico protein synthesis. *J. Comput. Chem.*, **26**, 1063–1068.

Peterson,R.W. *et al.* (2004) Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci.*, **13**, 735–751.

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

Rohl,C.A. *et al.* (2004) Protein structure prediction using rosetta. *Method Enzymol.*, **383**, 66–93.

Samudrala,R. and Moult,J. (1998) A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol.*, **279**, 287–302.

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

Tarjan,R. (1972) Depth-first search and linear graph algorithms. *SIAM J. Comput.*, **1**, 146–160.

Tuffery,P. *et al.* (1991) A new approach to the rapid-determination of protein side-chain conformations. *J. Biomol. Struct. Dyn.*, **8**, 1267–1289.

Vasquez,M. (1995) An evaluation of discrete and continuum search techniques for conformational-analysis of side-chains in proteins. *Biopolymers*, **36**, 53–70.

Wang,G.L. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Xiang,Z.X. and Honig,B. (2001) Extending the accuracy limits of prediction for side-chain conformations (vol 311, pg 421, 2001). *J. Mol. Biol.*, **312**, 419–419.

Xu,J.B. (2005) Rapid protein side-chain packing via tree decomposition. *Lect. Notes Comput. Sci.*, **3500**, 423–439.