doi:10.1006/jmbi.2001.4865 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 311, 421–430

JMB

Extending the Accuracy Limits of Prediction for Side-chain Conformations

Zhexin Xiang and Barry Honig*

Department of Biochemistry and Molecular Biophysics BB221 Columbia University Box 36, New York, NY 10032, USA

*Corresponding author

Current techniques for the prediction of side-chain conformations on a fixed backbone have an accuracy limit of about 1.0-1.5 Å rmsd for core residues. We have carried out a detailed and systematic analysis of the factors that influence the prediction of side-chain conformation and, on this basis, have succeeded in extending the limits of side-chain prediction for core residues to about 0.7 Å rmsd from native, and 94% and 89% of χ_1 and χ_{1+2} dihedral angles correctly predicted to within 20° of native, respectively. These results are obtained using a force-field that accounts for only van der Waals interactions and torsional potentials. Prediction accuracy is strongly dependent on the rotamer library used. That is, a complete and detailed rotamer library is essential. The greatest accuracy was obtained with an extensive rotamer library, containing over 7560 members, in which bond lengths and bond angles were taken from the database rather than simply assuming idealized values. Perhaps the most surprising finding is that the combinatorial problem normally associated with the prediction of the side-chain conformation does not appear to be important. This conclusion is based on the fact that the prediction of the conformation of a single side-chain with all others fixed in their native conformations is only slightly more accurate than the simultaneous prediction of all side-chain dihedral angles.

Keywords: side-chain prediction; rotamer library; protein structure; combinatorial property; force field

Introduction

Accurate prediction of side-chain conformation is an important element in homology modeling, in protein design and in flexible ligand docking.[1–3] Prediction accuracy depends on the effectiveness of the conformational search procedure and on the quality of the force-field used to evaluate the conformational energy.[4] Due to the combinatorial problem that arises from the fact that there is generally more than one conformation for each side-chain, exploring all possible combinations of side-chain conformations in a particular protein is computationally intractable. This problem is further complicated by the uncertainties associated with the calculation of the conformational free energies of proteins in solution. Indeed, the two problems are related and it has been difficult to determine if limitations in a particular methodology are due to inadequate sampling or to an inaccurate calculation of the conformational energy. Here, we demonstrate that conformational sampling is not a major problem and that even crude energy functions are sufficient to yield accurate predictions for buried side-chains.

Rotamer libraries, first introduced by Ponder & Richards,[5] have been widely used in the prediction of side-chain conformations. The Ponder & Richards library consisted of 67 rotamers that represented most observed side-chain conformations in subset of 19 well-refined proteins. Later work expanded the number of rotamers (see for example, Tuffery *et al.*[6] and Maeyer *et al.*[7]), and introduced backbone-dependent rotamer libraries (see, for example, Dunbrack & Karplus[8] and Bower *et al.*[9]). The use of a rotamer library reduces the conformational search problem and reduces the dependence of the result on the energy function. This is because the most frequently observed rotamers tend to be energetically favored.[10,11] An underlying problem with the use of rotamers is that there are many side-chains whose conformation is not close to one of the conformations

Abbreviations used: B&T, branch-and-terminate; PDB, Protein Data Bank.

E-mail address of the corresponding author: bh6@columbia.edu

0022-2836/01/020421–10 $35.00/0

included in the rotamer library.[11] This problem can be addressed, as discussed below, by using a very detailed rotamer library.

A number of approaches have been applied to the conformational search problem. These include simulated annealing,[12–14] genetic algorithms,[15,16] Monte Carlo calculations,[1,17] dead-end elimination,[2,18–22] segment matching,[23] molecular dynamics refinement[24] and mean field optimization.[17,25,26] Most recently, the ''Branch-and-Terminate'' method (B&T)[27] was used in the search of a combinatorial tree that represented a protein. In some applications, the B&T method was found to be 21 times faster than dead-end elimination. However, there is always a tradeoff between accuracy and speed (see, for example, Voigt *et al.*[28]).

Most side-chain prediction methods yield very similar results. This is perhaps not surprising, since all methods are, in many ways, closely related. They differ from one another primarily in two aspects; the selection of an initial conformation and the survival criteria used to produce the next generation in an iteration procedure. Most current methods yield rmsd values from native of between 1.0 and 1.5 core residues.[3,29] Expressed in terms of torsional angle, the percentage of of χ_1 and χ_{1+2} correctly predicted to within 20° is about 80% and 70%, respectively.[3] Prediction accuracy decreases significantly when all residues are considered. Recently, Samudrala & Moult[4] used an effective statistical potential to obtain an accuracy of 1.72 Å rmsd for all residues in ten proteins that were studied.

The accuracy achieved for core residues is close to the 1.0 Å rmsd limit suggested by Levitt *et al.*[4] based on comparisons of different crystal structures of the same protein. However, by fixing all residues, except the one being considered, in their native conformation, Petrella *et al.*[30] suggested a somewhat lower limit, of 0.8 Å rmsd. Here, we carry out a detailed study of the various factors that influence the accuracy of side-chain prediction so as to determine the maximum accuracy obtainable under different conditions. On the basis of our findings, we develop a new method for predicting side-chain conformations. For core residues, we are able to obtain accuracies of about 0.7 Å rmsd or, expressed in terms of the percentage of χ_1 and χ_{1+2} values correctly predicted within 20° from native, accuracies of about 94% and 89%, respectively. This level of accuracy is obtainable only when an extensive rotamer library, consisting of over 7000 distinct conformations, is used. However, the accuracy is dependent on the quality of the protein backbone. On the basis of our results, we argue that there is effectively no combinatorial problem associated with the prediction of buried side-chain dihedral angles and that essentially any method that uses a detailed rotamer library can succeed.

Results and Discussion

Evaluation of rotamer libraries

Side-chain prediction accuracy depends heavily on the rotamer library used. Current libraries are often incomplete and too coarse grained. Figure 1 plots the number of rotamers as a function of the number of proteins used to construct the library, for three levels of rotamer resolution. In all cases, the number of rotamers increases with the number of proteins but, as might be expected, the slope depends on the dihedral angle tolerance used to define a rotamer. In all, 75% of the rotamers come from only four amino acids, Glu, Gln, Lys and Arg, since they have more torsional degrees of freedom than the other residues. Met, which like Glu and Gln has three degrees of rotational freedom, has half the number of rotamers, perhaps due to its lower abundancy in proteins. Figure 2 plots the percentage of rotamers needed to represent a given percentage of all rotamers in a library *versus* the percentage represented. As an example, for the 10° library obtained from a set of 297 proteins (Figure 1), 62% of the rotamers ($62\% \times 7562 = 4696$) can represent 86% of the residues in the database from which the rotamer library was created; for 40° rotamer library, only 24% of the rotamers ($24\% \times 994 = 241$) are needed. These results are consistent with the fact that side-chain dihedral angles tend to cluster around energy minima corresponding to those of the isolated residues.

We used the first test set of 15 proteins to determine the ability of a given rotamer library to represent proteins not included in the creation of the library. The optimal representation of a particular residue is obtained by choosing the rotamer with the lowest rmsd from the native conformation. As evident from Figure 3, the accuracy of a rotamer library is only weakly dependent on the number of proteins in the library. Even though, for 20° resolution, there are almost six times the number of rotamers in the 2312 protein library than in the 135

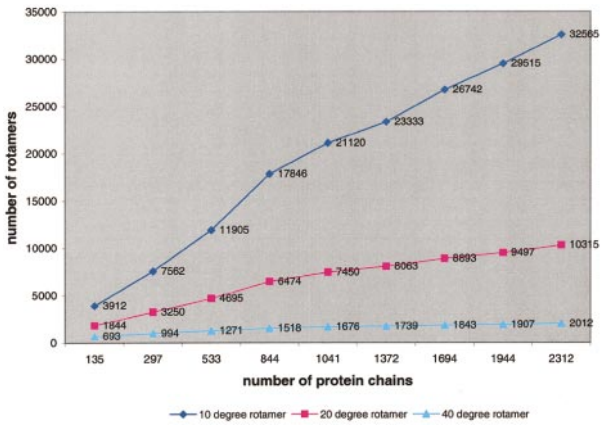

Figure 1. Number of rotamers plotted as a function of the nunber of proteins in the database.

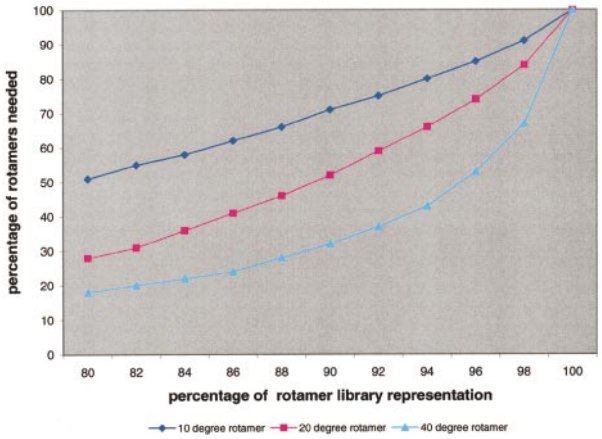

Figure 2. Number of rotamers needed to represent the side-chains present in the 297 protein database plotted against the percentage of the rotamers in the database.

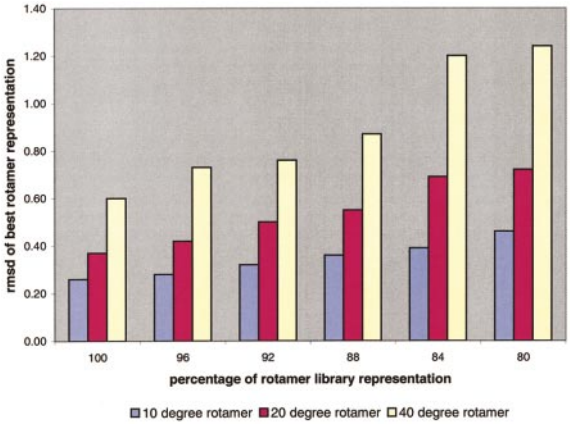

Figure 4. The rmsd from native of the best rotamer representation for rotamer libraries with differening percentage representation. The rotamer libraries were compiled from the database of 297 chains.

protein library (1844 *versus* 10315; see Figure 1), the rmsd of their best representation of the test set is not that different, 0.41 Å *versus* 0.36 Å (Figure 3). (The small maximum for the 1372 protein library is probably due to the removal by chance, of several rotamers that were particularly well matched to certain residues in the 15 protein test set). In contrast, the quality of the representation is strongly dependent on the resolution of the library that is used. As shown in Figure 4, using rotamers compiled from 297 proteins, 10° rotamers can represent the 15 proteins with an accuracy of 0.26 Å, compared to 0.6 Å for 40° rotamers. The completeness of the library becomes increasingly important as the resolution of the library decreases. For the 40° rotamer library, the rmsd of its best representation of the test set increases from 0.6 Å to 1.2 Å when the percentage of its representation of the original set of 297 proteins decreases from 100% to 84%; the increase is just from 0.2 7 Å to 0.39 Å for the 10° library.

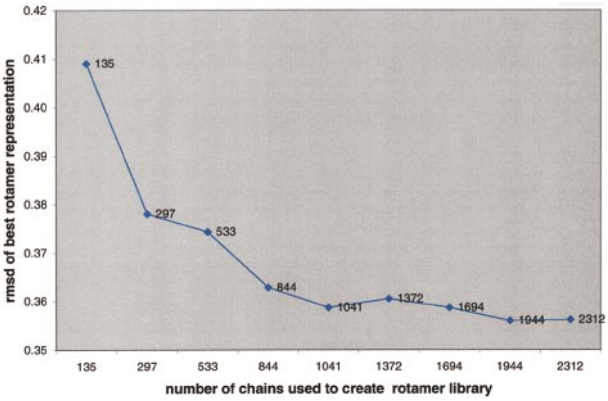

Figure 3. The rmsd of the best rotamer representation from the native for different databases. The numbers shown above the curve are the number of chains in the database from which the rotamer library was derived. Results are for a complete 20° rotamer library.

Given these results, it appears most practical, if CPU time is a concern, to use a high-resolution library taken from a limited set of proteins. In the following, we have used a library that represents 100% of the rotamers in the database of 297 proteins, which includes 7562 rotamers in total.

Accuracy limit of side-chain prediction

Prediction accuracy depends on several factors, such as the choice of force-field, combinatorial complexities, quality of the rotamer library, quality of the protein backbone and bond angle and length parameters. We have studied each of these factors separately in order to determine their effect on prediction accuracy. All the following studies were done on the first set of proteins listed in Table 4.

Effect of force-field

In this series of tests, bond lengths and angles were set to their experimental values in order to remove any effect of uncertainties in these parameters. The combinatorial problem was eliminated by predicting side-chain conformations, one residue at a time, while all others were held fixed in their native conformation. The native conformation was added to the rotamer library for each residue, so as to eliminate effects arising from the quality of the library. The only remaining factor is the quality of the force-field, in particular the representation of torsional potentials, van der Waals forces and electrostatic interactions.

The CHARMM[31] and AMBER[32] force-fields were used, each with a distance-dependent dielectric constant. Table 1 suggests that for these force-fields, including only van der Waals and torsion energy terms yields the best results. The inclusion

Table 1. Side-chain prediction accuracy using different force-fields

	CHARMM22				AMBER96			
	Core		All		Core		All	
	rmsd	χ_1/χ_{1+2}	*a*	*b*	*a*	*b*	*a*	*b*
vdw	0.55	92/87	1.75	73/59	0.64	91/84	1.75	72/56
vdw + tor	0.38	97/95	1.25	88/76	0.44	96/92	1.39	86/72
vdw + tor + crg	0.56	95/89	1.50	81/65	0.64	92/84	1.51	79/61

Side-chain predictions were performed for one side-chain each time with all others fixed at the native conformations. The native conformation is included in the rotamer library. vdw denotes the van der Waals interactions, intrinsic torsional energy, crg denotes charge-charge interactions with a distance-dependent dielectric constant. *a, rmsd; b*, percentage of χ_1/χ_{1+2} correctly predicted to within 20°; core implies residues that are more than 90% buried. A 10° rotamer library sampled from 297 proteins with 100% representation was used to obtain the results shown.

of partial charges decreases prediction accuracy, since solvent effects have not been accounted for. Including solvent effects using continuum methods has not been found to yield a significant improvement in prediction accuracy.[29,33] The explicit inclusions of individual water molecules does improve accuracy, primarily due to a more realistic representation of packing effects.[30] It is likely that an improved representation of solvent effects is important for surface residues. However, given the corresponding cost in CPU, we have simply ignored electrostatic interactions, an approximation that seems particularly well-suited for core residues. It appears then, that if solvent effects are not taken into account, the most accurate prediction of buried side-chain conformations is obtained when only van der Waals and torsional terms are used to evaluate conformational energies. Most previous work did not include intrinsic dihedral energies in side-chain predictions.

In order to determine if the approximately 0.1 Å rmsd preference of CHARMM with respect to AMBER (Table 2) is due to the ten pre-minimization steps applied to the original coordinates using CHARMM, we repeated the calculations using non-minimized structures. The results were very similar, with CHARMM yielding rmsd values of about 0.40 Å *versus* 0.46 Å for AMBER. This means that CHARMM is slightly better than AMBER in our side-chain prediction. All results given below are obtained using CHARMM. Table 2 suggests that, under optimal conditions, side-chain prediction accuracy for core residues can be as low as 0.38 Å rmsd with 97% and 95% of the dihedral angles predicted to within 20° for χ_1 and χ_{1+2}, respectively. The most severe approximation made in obtaining these results is clearly the restriction of all side-chains but one to their native conformations, which removes the combinatorial complexity of the problem. The effect of this approximation is explored in the following section.

Table 2. Side-chain prediction accuracy obtained using the best rotamer initial conformations

	Core		All	
	rmsd	χ_1/χ_{1+2}	rmsd	χ_1/χ_{1+2}
(a)	0.39	97/95	1.30	88/76
(b)	0.47	97/92	1.32	87/74

(a) Native conformation included as rotamer; (b) native conformation not included.

The effect of the combinatorial problem

Since side-chain prediction is an N-body interaction problem, finding the correct global energy minimum is NP-hard and is thus impossible.[4,30] Here, we will demonstrate that, in practice, the combinatorial problem poses no significant limitations on prediction accuracy. In the previous section, the combinatorial problem was removed by placing all side-chains but the one being tested in their native conformation.

We first test the effects of removing the native conformation as one of the rotamers for the side-chain being tested. The calculations reported in Table 2 were repeated but without fixing any side-chains in their native conformations. The native conformation of each residue was retained in the rotamer library but the initial conformation chosen for each residue was the closest rotamer to native rather than the native conformation itself. Predictions were then carried out, as described above, one residue at a time until convergence is obtained. As can be seen in Table 2 (row (a)), the accuracy obtained in this way is 0.39 Å, close to the value of 0.38 Å reported in Table 1. This indicates that only trivial errors are introduced when rotamers are not fixed at their native conformations. The next logical step is to remove the native conformation entirely from the rotamer library.

Row (b) of Table 2 is obtained from a set of calculations identical with that used to produce row (a) but in this case the native conformation is not included in the rotamer library. This results in a decrease in accuracy of 0.08 Å rmsd relative to the case where the native conformation is included. Row (b) of Table 2 provides a realistic estimate of the most accurate results possible if the native conformation is not known. That is, the best side-chain prediction accuracy obtainable from current force-fields is about 0.47 Å rmsd, 97% correct for χ_1 and 92% correct for χ_{1+2}. In the following paragraph we examine if these limits can still be obtained

when the best initial conformation is not known in advance.

The lowest-energy conformation obtained in the 120 minimizations was used as the prediction for a given protein. As shown in row a of Table 3, the prediction accuracy for core residues is 0.51 Å rmsd, with 96 % and 92 % correct to within 20° for χ_1 and χ_{1+2}, respectively. Comparing row (a) of Table 4 and row (b) of Table 3, it appears that an arbitrary choice of starting conformation lowers the prediction accuracy (relative to the optimal choice) by only 0.04 Å for core residues and 0.14 Å for all residues. This suggests that there is essentially no combinatorial problem in side-chain prediction.

The accuracy of 0.51 Å obtained in Table 3, row (a), is close to what can be obtained in real applications. This is because the only simplification that has been used involves the use of side-chain native bond angles and lengths. The effect of this approximation will be explored in the next section. The overall prediction accuracy depends critically on the degree of burial, chemical nature and secondary structure assignment of the residue. As can be seen in Figure 5, for totally buried residues, the rmsd is only 0.36 Å; when the residue is more than 90 % exposed, the rmsd increases to 3.68 Å, no better than random.[3] Of course there are uncertainties associated with the experimental assignment of surface side-chain conformers as well.

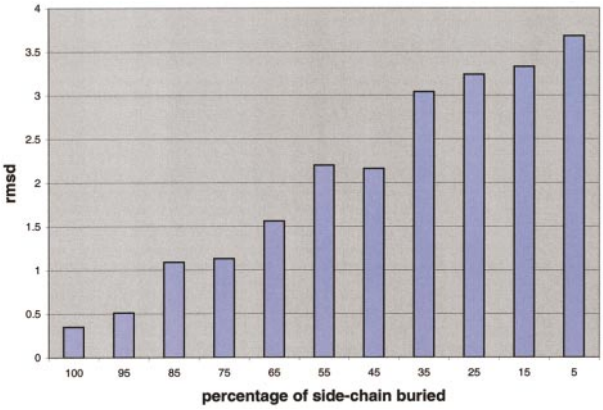

Figure 5. Side-chain prediction accuracy for residues with different percentage burial.

Side-chain bond angle and bond lengths

All the above calculations were performed using native bond angles and bond lengths for side-chains. In real applications these will not be known; however, their effect is significant. For example, if native dihedral angles are used to reconstruct side-chain conformation based on the standard bond angles and bond lengths in CHARMM, the rmsd from native is as large as 0.26 Å (data not shown). The calculations leading to the results in row (a) of Table 3 were repeated using bond angles and lengths taken from CHARMM. The results are reported in row (b) of Table 3. Comparison of rows (b) and (a) in this Table reveals an accuracy decrease for core residues of about 0.53 Å rmsd and decreases of about 5 % and 9 % for correctly predicted χ_1 and χ_{1+2} values, respectively.

Interestingly, if we use the coordinate rotamer library, the results are significantly improved as shown in row (c). In this case, the accuracy decrease relative to row (a) is just 2 % and 5 % in χ_1 and χ_{1+2}, and 0.29 Å in rmsd. Remarkably, the effect of not knowing side-chain bond lengths and bond angles is more severe than that of the combinatorial problem. This is not difficult to understand, because the coordinate rotamer library accounts, in part, for correlations in bond lengths and bond angles with dihedral angles that are not present in the standard rotamer library. In this regard, it has been shown recently that all published detailed rotamer libraries contain some rotamers that exhibit unacceptable steric clashes if used with idealized geometries and explicit hydrogen atoms.[34] The total energy of a protein conformation predicted using coordinate rotamers is lower than that using dihedral rotamers. The advantage of using coordinate rotamers decreases if a 40° rotamer library is used.

Table 3. Side-chain prediction accuracy using 120 different initial conformations

	Core		All	
	rmsd	χ_1/χ_{1+2}	rmsd	χ_1/χ_{1+2}
(a)	0.51	96/92	1.46	85/72
(b)	1.04	91/83	1.67	81/62
(c)	0.80	94/87	1.59	83/65
(d)	0.62	95/91	1.49	85/68
(e)	0.71	94/89	1.56	84/66
(f)	1.22	90/69	1.97	78/53
(g)	1.37	88/68	1.78	81/56

(a) Native side-chain bond angles and lengths.
(b) CHARMM's bond lengths and angles.
(c) Coordinate rotamer library.
(d) Structure refinement performed on row (b).
(e) No pre-minimization, proline conformation predicted.
(f) Backbone-dependent rotamers.
(g) Results from Dunbrack's SCWRL program.

Structure refinement

In order to determine if the 0.8 Å result given in row (b) of Table 3 can be improved, we carried out a simple structure refinement for each of the 120 predictions made for each protein. The refinement simply involved sampling conformations that involved ±2° rotations of each dihedral angle for each residue. The procedure was carried out, one residue at a time, with all other residues kept fixed. The entire process was terminated when further iterations produced no change in conformation. As can be seen in row (d) of Table 3, the accuracy is improved to 0.62 Å in rmsd for core

Table 4. Side-chain prediction accuracy for the first test set

PDB	Res.	*R*	rmsd0	Core rmsd	Core χ_1/χ_{1+2}	All rmsd	All χ_1/χ_{1+2}
1cbn	46	0.83	0.051	0.44	100/100	1.45	92/69
1cex	214	1	0.057	0.67	98/96	1.48	85/82
5pti	58	1	0.091	0.2	100/100	1.55	83/74
1ixh	321	0.98	0.061	0.44	92/92	1.34	83/71
2pth	193	1.2	0.052	0.5	95/91	1.55	86/68
5p21	166	1.35	0.068	1.22	89/79	1.82	78/59
1aho	64	0.96	0.07	0.7	90/90	1.83	88/76
3lzt	129	0.92	0.07	0.45	100/92	1.59	89/68
1ctj	89	1.1	0.057	0.7	95/88	1.45	84/67
1igd	61	1.1	0.066	0.15	100/85	1.34	84/61
7rsa	124	1.26	0.067	0.5	97/85	1.91	81/55
1aac	105	1.31	0.062	0.43	97/82	1.12	94/67
1eca	136	1.4	0.079	0.31	98/94	1.3	86/70
1plc	99	1.33	0.072	0.36	95/93	1.14	83/70
1rcf	169	1.4	0.058	0.68	97/93	1.3	84/67

R is the structure resolution in Å. rmsd0 relates to the difference between CHARMM22 ten-step pre-minimized structure and the original structure. rmsd is the difference between the predicted side-chain conformation and the native. Dihedral angles within 20° of the native value are considered to be correct. Res. is the number of residues in the target.

residues and by 1-3% for χ_1 and χ_{1+2}, respectively.

The results in Table 3 row (d) were obtained by comparing the predicted side-chain conformation with pre-minimized structure instead of the original one. The effect of comparing to the raw Protein Data Bank coordinates is to increase the average rmsd for core residues by 0.003 Å, while the average rmsd for all residues decreases by 0.002 Å. We also carried out calculations where no pre-minimization was applied and where proline conformations were predicted rather than kept fixed. As can be seen in Table 4 row (e), the rmsd for core residues increases slightly , to 0.71 Å. Table 4 shows results for each protein in the test set; prediction accuracy for core residues ranges from 0.15 Å in 1igd to 1.22 Å in 5p21.

Effects of rotamer library

The improvement in prediction accuracy obtained here relative to previous work results comes mainly from our use of an improved rotamer library. In order to illustrate this point, we repeated our calculations using the same energy function, search scheme, and definition of core residues but using Dunbrack's backbone-dependent rotamer library. The 1.22 Å rmsd obtained for core residues as shown in Table 3 row (f) is less accurate than we obtained with a more extensive rotamer library. Under the same evaluation criteria, we also repeated predictions with Dunbrack's SCWRL program.[31] The rmsd accuracy for core residues is 1.37 Å (row (g)), which is similar to that listed in row (f). The reason that our calculation using Dunbrack's rotamer library is slightly more accurate than SCWRL is that we searched conformation space more extensively.

Prediction results on the second test set

In the first test set, the prediction accuracy obtained for core residues was 0.71 Å rmsd. In order to remove any possible bias that might have been introduced from extensive testing on this set, a second test set of proteins was chosen as described in Materials and Methods. No pre-minimization was carried out for this set, and predictions were made for all proline residues. As shown in Table 5, the average prediction accuracy obtained for core residues is 0.74 Å rmsd and 95% and 86% for χ_1 and χ_{1+2}, respectively. If agreement with experimetnt is defined as being within 40° of the native conformation, the accuracy for core residues improves to 96% and 90% for χ_1 and χ_{1+2}, respectively. These results are very similar in accuracy to those obtained for the first test set.

Conclusion

We have carried out a systematic study of the factors that determine the accuracy of the prediction of side-chain conformation. A central result is that prediction accuracy depends primarily on the quality of the rotamer library that is used. The essential elements in a good library include high resolution and the use of different bond lengths and bond angles for each rotamer in the library. We found that the use of a backbone-dependent rotamer library actually degrades results, since it places unnecessary restrictions on the number of rotamers that are used. The results obtained in this work were based on a simple conformational search method that could be improved easily with more advanced methods reported in the literature.[27] Nevertheless, predictions on a typical protein take only about two hours on an R10000 workstation. We are in the process of developing

Table 5. Side-chain prediction accuracy for the second test set

PDB	Res.	*R*	Core rmsd	Core χ_1/χ_{1+2}	All rmsd	All χ_1/χ_{1+2}
1b9o	123	1.15	0.73	95/81	1.7	83/60
1c5e	95	1.1	0.49	92/88	1.37	77/66
1c9o	66	1.17	0.44	100/83	2.15	77/60
1cc7	73	1.2	0.18	100/90	1.57	85/66
1cku	85	1.2	0.33	100/83	1.16	92/76
1cz9	162	1.2	0.64	100/88	1.86	91/74
1czp	98	1.17	0.81	92/87	1.41	84/67
1d4t	104	1.1	0.9	88/75	1.99	79/61
1mfm	153	1.02	0.49	98/82	2.07	75/56
1qj4	257	1.1	0.65	96/91	1.27	92/80
1qnj	240	1.1	1.03	90/80	1.65	79/60
1ql0	241	1.1	0.57	97/87	1.33	89/70
1qlw	328	1.1	0.63	94/85	1.6	85/72
1qtn	164	1.2	0.55	95/91	2.01	80/60
1qtw	285	1.02	0.67	96/89	1.64	85/70
1qu9	128	1.2	1.31	87/74	1.83	72/54
1vfy	73	1.15	0.25	100/100	1.57	81/60
1qq4	198	1.2	0.59	95/92	1.32	86/79
AVG			0.74	95/86	1.66	84/67

Symbols have the same meaning as in Table 4.

an energy-mapping, grid-based technique that appears to result in speed-ups of one to two orders of magnitude.

Perhaps the most surprising finding of this work is that there does not appear to be a significant combinatorial problem in side-chain prediction. In fact, although 120 initial conformations were chosen for each protein, about 10% of these resulted in conformations with rmsd to the native similar to that (within 0.2 Å difference) of the lowest energy conformation. Thus, one could reduce the number of starting conformations by almost an order of magnitude and still obtain comparable results with a corresponding reduction in CPU time. The apparent lack of combinatorial complexity may be due, in part, to restrictions placed by the main chain on side-chain conformations. However, side-chain predictions based on main chain alone are less accurate than those that address the full combinatorial problem.[35] Thus, as pointed out by Levitt *et al.*,[4] combinatorial packing is an important feature of the side-chain conformation prediction problem. However, as shown here, the combinatorial problem can be overcome easily without large CPU cost.

For real applications, we have found that a prediction accuracy of 0.7 Å for core residues is easily obtainable. This result might even be improved somewhat in homology modeling applications, where rotamer preferences from one protein might be transferable to another. Nevertheless, the accuracy obtained in this work for core residues is close to the limits imposed by experimental error, and intrinsic protein motion, so that in this sense this part of the problem may perhaps be viewed as solved. This is far from the case for surface residues, where improvements in energy functions and solvation models are clearly necessary.

A second part of the problem is that side-chain prediction accuracy depends on the accuracy with which the backbone is represented.[16,36] We have found similar results in our calculations on CASP4 targets (unpublished results). Since in many comparative modeling applications the backbone coordinates are not known accurately, the results obtained here are unlikely to be duplicated in many real problems. However, this issue can be thought of as separable from the side-chain prediction problem for cases where the backbone conformation is known. In fact, the results obtained here suggest that an approach to refinement of homology models through alternate cycles of side-chain and backbone optimization[23] may prove fruitful.

Materials and Methods

Selection and preparation of rotamer libraries

Test set of proteins

Two sets of proteins solved to high resolution were selected from the Protein Data Bank (PDB) and used as a test of our approach. The first set included 15 proteins selected randomly from the culled PDB list of Dunbrack in 1998 with resolution between 0.83-1.4 Å. All proteins in this list had a pair-wise sequence identity of less than 20%. The second set included all proteins deposited in the PDB in 1999 with resolution better than 1.2 Å and that contained more than 40 residues. For multi-chain proteins, we used only the first chain. The second test contains 18 proteins. Hydrogen atoms were added using What if.[37] The first set of proteins was used to evaluate the influence of different parameters on the accuracy of side-chain prediction, while the second set was used to test whether the parameters used for the first set were also able to produce good results on an independent set of proteins.

The conformational energies of the first set of proteins were minimized with ten steps of steepest descent using the CHARMM (PARAM22) force-field.[38] The rmsd

between the minimized and native structures was quite small, usually less than 0.07 Å (see Table 4). The purpose of the minimization is to remove steric clashes that are present in the crystal structures (see, for example, Petrella *et al*.[30]). We will demonstrate below that the pre-minimization has no significant influence in the side-chain prediction.

Proteins used for the rotamer library

Rotamers were generated from nine non-redundant databases created by Dunbrack in 1998 that contain different numbers of proteins 135, 297, 533, 844, 1041, 1372, 1694, 1944, 2312 proteins, respectively†. All proteins in the test sets listed in Tables 4 and 5 were removed from the databases used to generate rotamers. Hydrogen atoms were added to the remaining proteins with the WhatIf program[37] and hydrogen atom coordinates were subjected to energy optimization using the CHARMM22 force-field.

Creation of rotamer library

For each database, we created three (10, 20, 40°) rotamer libraries. As an example, in the creation of a 10° rotamer library, two side-chains are considered to have equivalent conformations if all of their corresponding side-chain torsion angles are within 10° of one other. Thus, the second side-chain will be represented by the first and will not be used to generate a new rotamer. Only the first four torsional angles were used to create rotamers for arginine residues but, unlike previous rotamer libraries where the last torsion angle was assigned as 180°,[5,6,8] we retained the native values for this angle in each rotamer in the library. This led to somewhat improved predictions, particularly for arginine residues.

Some rotamers occur more often than others and some rotamers occur very infrequently. Generally, a small portion of a complete rotamer library can represent a large fraction of the side-chain conformations that are observed in a set of proteins. A 90% rotamer library is defined as representing 90% of the conformations in the database used to create it. In order to create such a library, the following procedure is applied to each of the 20 amino acids. Using valine as an example, if there are a total number of 100 distinct rotamers in the complete rotamer library, and if the total number of valine residues in the database is 10000, we take the smallest number of rotamers from the complete rotamer library that can represent 90% (9000) of the valine residues in the database.

In addition to representing rotamers in terms of dihedral angles, as is normally done, we represented rotamers in terms of Cartesian coordinates, thus maintaining bond lengths and angles at their experimental values for a given rotamer. We designate the two kinds of libraries, dihedral angle and coordinate libraries, respectively. As will be discussed below, the latter yields more accurate predictions. As pointed out by a referee, the rotamers we have selected do not necessarily lie at the bottom of local energy wells and it is worth distinguishing the usage of the term used here and elsewhere from the one that implies a local energy minimum.

† http://www.fccc.edu/research/labs/dunbrack/index.html

Side-chain prediction method

Side-chain predictions were carried out on fixed polypeptide backbones, which includes N, C^{α}, C, O and C^{β} atoms and their associated protons. We considered C^{β} as part of the backbone because its position can be determined uniquely given the coordinates of N, C^{α} and C. All the atoms on the backbone were fixed, except hydrogen atoms on C^{β}, which must rotate with the side-chains. No prediction was made for Ala or Gly. The side-chain positions of cysteine residues forming disulfide bonds were generated on the basis of geometric considerations. In the first series of calculations, proline residues were kept fixed in their native conformations. As will be shown below, this simplification has only marginal effects on the results. For each residue, all possible side-chain conformations are generated either from the dihedral angle rotamer library using bond angles and lengths from CHARMM, or from their coordinate rotamer library.

Given an initial conformation, energy minimization is performed one residue at a time, while all other residues are kept fixed. The minimization procedure simply tests all rotamers for a given residue and picks the one that is lowest in energy. The minimization proceeds from the first residue to the last, and the procedure is repeated until all the side-chain conformations retain the same rotamer upon further iteration. Some rotamers can be removed from consideration during this procedure. For each residue, we ignored rotamers with over 100 kcal/mol (1 cal = 4.184 J) interaction energy with the backbone. If more than 100 rotamers pass this filter, the first 100 with the lowest interaction energy with the backbone are retained. A total of 120 different initial conformations are used for each protein. The first calculation for each protein begins with a conformation where each rotamer is at its lowest energy with respect to the backbone. The next 59 calculations use totally random initial conformations. The remaining 60 calculations use initial conformations that are obtained from a simple partial randomization procedure applied to the lowest-energy conformation obtained from the 60 previous runs in which side-chains of negative or positive interaction energy have 30% or 70% probabilities, respectively, of being replaced by a random rotamer. The set of side-chain conformations assigned to a particular protein is chosen as the lowest-energy conformation obtained from the 120 sets of calculations. In some cases, a further refinement procedure was carried out as described below.

The program, SCAP, is available at honiglab.cpmc.columbia.edu.

Accuracy analysis

The accuracy of side-chain conformer predictions is usually assessed in terms of rmsd values from the native conformation, or in terms of dihedral angle deviations from native. As the two criteria are not consistent with one another,[8] most of the time, both are used in this work. In the calculation of rmsd, we considered only side-chain heavy atoms, excluding C^{β} atoms. There are two ways to calculate rmsd among several proteins: overall rmsd *versus* average rmsd. For example, in the case of two proteins, the overall rmsd is calculated by summing over all residues in both proteins while the average rmsd is simply one half of the sum of rmsd for each of the two proteins. The value of overall rmsd is

usually larger than that of average rmsd and is used here.

Only χ_1 and χ_{1+2} dihedral angles were considered when accuracy was measured in terms of side-chain dihedral angles. Other side-chain dihedral angles are not evaluated, though they were allowed to rotate in the prediction procedure. A dihedral angle was considered to be predicted correctly if its value was within 20° of that of the minimized native structure. The average side-chain dihedral angle difference between the minimized and native structure is less than 1°. In evaluating the results, the symmetry in Asp, Glu, Phe, Tyr and Arg residues was taken into account, as was the equivalence in the two terminal oxygen atoms in Asp and Glu residues and the NH1 and NH2 atoms in arginine residues.

Prediction accuracy is correlated with the degree of residue burial, measured in terms of the percentage of the solvent-accessible area of the side-chain relative to that of an isolated side-chain in an extended conformation. Here, protein core residues are defined as residues that are more than 90% buried. Definitions of core residues in the literature have varied between 70 and 90% burial. For example, Dunbrack & Karplus,[8] Shenken *et al.*,[14] and Maeyer *et al.*[7] used 90% bural to define core, while Holm & Śander (1991) used 80%. The calculated percentage burial is fairly sensitive to the description of the amino acid in a random peptide. Holm & Sander[1] defined the conformation of an individual amino acid by placing it in an extended Gly-Gly-X-Gly-Gly peptide. On the basis of this definition, about 40% of the residues in 20 proteins were labeled as core. Dunbrack & Karplus[8] used the peptide acetyl-X-$NHCH_3$ as a reference state. Our reference state is that of an isolated amino acid that has a larger solvent-accessible surface area than an extended peptide. In fact, although we use 90% burial to define our core residues, our use of an isolated amino acid as a reference state results in a larger percentage of residues labeled as core (45%) than were labeled by Holm & Sander.[1]

Acknowledgements

This work was supported by NIH grant GM-30518. We are grateful to Drs A. Yang, J. Norberg, J. Jong and F. Sheinerman for stimulating discussions.

References

1. Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain coordinates from CA trace: application to model building and detection of coodinate errors. *J. Mol. Biol.* **218**, 183-194.
2. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA,* **94**, 10172-10177.
3. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* **66**, 549-579.
4. Samudrala, R. & Moult, J. (1998). Determinants of side-chain conformational preferences in protein structures. *Protein Eng.* **11**, 991-997.
5. Ponder, J. W. & Richard, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumation of allowed sequence for different structure classes. *J. Mol. Biol.* **193**, 775-791.
6. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267-1289.
7. Maeyer, M. D., Desmet, J. & Lasters, I. (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.
8. Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.* **230**, 543-574.
9. Bower, M., Cohen, F. E. & Dunbrack, R. L., Jr (1997). Homology modeling with a backbone-dependent rotamer library. *J. Mol. Biol.* **267**, 170-184.
10. Dunbrack, R. L., Jr & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Struct. Biol.* **1**, 334-340.
11. Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: to be or not to be? an analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **230**, 592-612.
12. Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373-388.
13. Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for side-chain optimization in proteins: application to model building by homology. *Proteins: Struct. Funct. Genet.* **14**, 213-223.
14. Shenkin, P. S., Farid, H. & Fetrow, J. S. (1996). Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins: Struct. Funct. Genet.* **26**, 323-352.
15. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1993). A critical comparision of search algorithm applied to the optimization of protein sidechain conformations. *J. Comput. Chem.* **14**, 790-798.
16. Tuffery, P., Etchebest, C. & Hazout, S. (1997). Prediction of protein side-chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng.* **10**, 361-372.
17. Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918-939.
18. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature,* **356**, 539-542.
19. Keller, D. A., Shibata, M., Marcus, E., Ornstein, R. L. & Rein, R. (1995). Finding the global minimum: a fuzzy end elimination implementation. *Protein Eng.* **8**, 893-904.
20. Lasters, I., Desmet, J. & De Maeyer, M. (1997). Dead-end based modeling tools to explore the sequence space that is compatible with a given scaffold. *J. Protein Chem.* **16**, 449-452.
21. Leach, A. R. & 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.
22. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429-445.

23. Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507-533.
24. Correa, P. E. (1990). The building of protein structures from alpha-carbon coordinates. *Proteins: Struct. Funct. Genet.* **7**, 366-377.
25. Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249-275.
26. Vásquez, M. (1995). An evaluation of discrete and continuum search techniques for conformational analysis of side-chains in proteins. *Biopolymers,* **36**, 53-70.
27. Gordon, D. B. & Mayo, S. L. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein desgin. *Struct. Fold. Des.* **7**, 1089-1098.
28. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789-803.
29. Vásquez, M. (1996). Modeling side-chain conformations. *Curr. Opin. Struct. Biol.* **6**, 217-221.
30. Petrella, R., Lazaridis, T. & Karplus, M. (1998). Protein sidechain conformer prediciton: a test of energy function. *Fold. Des.* **3**, 353-377.
31. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swamirathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamic calculations. *J. Comput. Chem.* **4**, 187.
32. Weiner, S. T., Kollman, P. A., Case, D. A., Singh, U., Ghio, C., Alagona, G. & Profeta, S. (1984). A new force field for molecular mechanism simulation of necleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765-784.
33. Wilson, C., Gregoret, L. M. & Agard, D. (1993). Modeling side-chain conformation for homologous protein using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996-1006.
34. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins: Struct. Funct. Genet.* **40**, 389-408.
35. Tanimura, R., Kidera, A. & Nakamura, H. (1994). Determinants of protein side-chain packing. *Protein Sci.* **3**, 2358-2365.
36. Huang, E. S., Koehl, P., Levitt, M., Pappu, R. V. & Ponder, J. W. (1998). Accuracy of side-chain prediction upon near-native protein backbones generated by Ab initio folding methods. *Proteins: Struct. Funct. Genet.* **33**, 204-217.
37. Vriend, G. (1990). WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56.
38. Mackerell, A. D., Jr, Bashford, D., Bellot, M., Dunbrack, R. L., Jr, Evanseck, J. D., Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. sect. B,* **102**, 3586-3616.

Edited by F. Cohen

(Received 22 September 2000; received in revised form 8 June 2001; accepted 12 June 2001)