J. Mol. Biol. (1997) **267**, 1268–1282

JMB

Prediction of Protein Side-chain Rotamers from a Backbone-dependent Rotamer Library: A New Homology Modeling Tool

Michael J. Bower[1], Fred E. Cohen[1,2] and Roland L. Dunbrack, Jr[1,2]*

[1]*Departments of Pharmaceutical Chemistry and* [2]*Cellular and Molecular Pharmacology University of California San Francisco, San Francisco, CA 94143-0450 USA*

**Corresponding author*

Modeling by homology is the most accurate computational method for translating an amino acid sequence into a protein structure. Homology modeling can be divided into two sub-problems, placing the polypeptide backbone and adding side-chains. We present a method for rapidly predicting the conformations of protein side-chains, starting from main-chain coordinates alone. The method involves using fewer than ten rotamers per residue from a backbone-dependent rotamer library and a search to remove steric conflicts. The method is initially tested on 299 high resolution crystal structures by rebuilding side-chains onto the experimentally determined backbone structures. A total of 77% of χ_1 and 66% of χ_{1+2} dihedral angles are predicted within 40° of their crystal structure values. We then tested the method on the entire database of known structures in the Protein Data Bank. The predictive accuracy of the algorithm was strongly correlated with the resolution of the structures. In an effort to simulate a realistic homology modeling problem, 9424 homology models were created using three different modeling strategies. For prediction purposes, pairs of structures were identified which shared between 30% and 90% sequence identity. One strategy results in 82% of χ_1 and 72% χ_{1+2} dihedral angles predicted within 40 degrees of the target crystal structure values, suggesting that movements of the backbone associated with this degree of sequence identity are not large enough to disrupt the predictive ability of our method for non-native backbones. These results compared favorably with existing methods over a comprehensive data set.

Keywords: protein structure; backbone-dependent rotamer library; side-chain prediction; homology modeling

Introduction

High quality models of protein structures for which we lack experimentally determined coordinates are necessary for many structure-based drug design efforts. However, solving the *de novo* protein folding problem for a protein sequence of interest has proven very difficult (Defay & Cohen, 1995). Modeling target structures by homology has been shown to be effective and useful in structure-based drug design (for reviews see Ring & Cohen, 1993; Bamborough & Cohen, 1996), and is particularly promising in light of the rapid growth of the sequence database compared to the database of known protein structures. In homology modeling, a model for a target protein is generated based upon the known structure of an homologous protein. Typically a model backbone is constructed for the structurally conserved regions, loops are added, and side-chains are placed (Browne *et al.*, 1969; Blundell *et al.*, 1987; Sutcliffe *et al.*, 1987a,b).

Our focus is the development of an algorithm for rapidly predicting side-chain conformation for homology modeling, where an amino acid sequence is built onto a protein backbone which may or may not be the native backbone for that sequence. Because of the interdependence of backbone and side-chain conformations, it is desirable to have a fast side-chain modeling tool that can be used with any number of backbone models. For example, we anticipate that this tool will be a useful adjunct for loop construction, one of the more difficult aspects of homology modeling. In ad-

Abbreviations used: PDB, Protein Data Bank; SCWRL, side-chains with a rotamer library.

0022–2836/97/151268–15 $25.00/0/mb970926

dition, predicting side-chain conformations from a sequence and a model of a protein backbone is a natural sub-problem of many current efforts in protein structure prediction and determination. On the theoretical side, these include *de novo* protein structure prediction inverse folding and threading algorithms, and protein folding simulations. In X-ray crystallography, this method could speed the initial placement of side-chains following the tracing of the main-chain through electron density maps prior to refinement calculations.

The combinatorial nature of side-chain placement on a given main-chain has often been cited as the main obstacle to predicting side-chain positions (Lee & Subbiah, 1991; Eisenmenger *et al.*, 1993). Historically, this has been addressed with a variety of strategies (Summers & Karplus, 1989; Lee & Subbiah, 1991; Tuffery *et al.*, 1991; Desmet *et al.*, 1992; Holm & Sander, 1992; Levitt, 1992; David, 1993; Dunbrack & Karplus, 1993; Eisenmenger *et al.*, 1993; Lasters & Desmet, 1993; Wilson *et al.*, 1993; Koehl & Delarue, 1994; Kono & Doi, 1994; Laughton, 1994; Hwang & Liao, 1995; Vasquez, 1995). Each effort in this field includes a decision about the conformational space each side-chain is allowed to sample, the energy function for evaluating solutions, and the choice of moves from one possible solution to the next; in other words, the rotamer set, the energy function, and the search strategy.

The observation that side-chains tend to exist in certain energetically favored conformations, or rotamers (Chandrasekaran & Ramachandran, 1970; Sasisekharan & Ponnuswamy, 1970; 1971; Janin *et al.*, 1978; Bhat *et al.*, 1979; Benedetti *et al.*, 1983; James & Sielecki, 1983; Ponder & Richards, 1987) has been used effectively to reduce the search space. Rotamer sets are usually derived from statistical analysis of experimental structures and consist of a list of conformations and their observed frequencies. In most cases, these conformations correspond to local minima on the side-chains' potential energy maps (Gelin & Karplus, 1979) and their relative probabilities correspond to what would be predicted from conformational analysis (Dunbrack & Karplus, 1994). Some authors use much larger sets of conformers derived from cluster analysis of the database of known structures. These often include very rare conformations, some of which must have very large dihedral strain energies (Schrauber *et al.*, 1993). Some rotamer sets have been developed which recognize a relationship between the side-chain conformation and the local backbone conformation (McGregor *et al.*, 1987; Dunbrack & Karplus, 1993).

For side-chain prediction, some authors have used rotamer sets which are independent of the local environment of the side-chain (Tuffery *et al.*, 1991; Holm & Sander, 1992; Koehl & Delarue, 1994), while others have used environment-dependent sets (Levitt, 1992; Dunbrack & Karplus, 1993; Eisenmenger *et al.*, 1993; Laughton, 1994). Other authors have augmented their set of rotamers with other conformers which offset low energy rotameric χ values by 10° or more (Desmet *et al.*, 1992; Tanimura *et al.*, 1994; Vasquez, 1995). Lee & Subbiah (1991) used a much broader, but still discrete, set of side-chain conformations in their work by incrementing χ angles 10° at a time. Other methods combine discrete conformers in the search strategy with a continuous minimization in the final stage of refinement (Dunbrack & Karplus, 1993; Vasquez, 1995). Larger rotamer sets have generally not proven more useful for side-chain predictions than smaller sets (Holm & Sander, 1992; Laughton, 1994; Tanimura *et al.*, 1994; Vasquez, 1995).

The fastest energy functions focus upon very local interactions and are typically limited to van der Waals or hard-sphere energies (Lee & Subbiah, 1991; Holm & Sander, 1992; Koehl & Delarue, 1994; Laughton, 1994; Vasquez, 1995). More detailed and longer-ranging energy functions slow the search process, and have not shown a significant improvement over these simple functions (Wilson *et al.*, 1993; Vasquez, 1996).

The different search strategies used have included Metropolis Monte Carlo methods (Holm & Sander, 1992), Gibbs sampling Monte Carlo (Vasquez, 1995), genetic algorithms (Tuffery *et al.*, 1991), neural networks (Hwang & Liao, 1995), simulated annealing (Lee & Subbiah, 1991; Hwang & Liao, 1995), mean-field optimization (Koehl & Delarue, 1994), the elimination of incompatible side-chain pairs (Desmet *et al.*, 1992; Lasters & Desmet, 1993), and ignoring combinatorial packing altogether (Eisenmenger *et al.*, 1993), as well as actual combinatorial searches (Tuffery *et al.*, 1991; Wilson *et al.*, 1993; Dunbrack & Karplus, 1993). In general, methods for predicting side-chain positions seem to be limited not by the quality of search algorithms, but rather by the quality of the energy functions employed and the approximations introduced by various rotamer sets (Tuffery *et al.*, 1993; Vasquez, 1996).

The method we describe here is based on the hypothesis that a great deal of the information needed for side-chain positioning is contained in the local main-chain conformation of each residue, but that a search strategy to resolve steric exclusions is also necessary for the most accurate predictions. This approach is implemented in the algorithm SCWRL, (side-chains with a rotamer library), which is a single self-contained program optimized for speed and accuracy. SCWRL is designed to take full advantage of the rotamer approximation and the strong backbone dependencies rotamers display (Dunbrack & Karplus, 1993; 1994) to create an initial placement for each residue, followed by systematic searches to resolve steric clashes.

Most previous methods for side-chain prediction have been tested only on side-chains built upon their native backbone (Tuffery *et al.*, 1991; Levitt,

1992; Laughton, 1994; Hwang & Liao, 1995; Vasquez, 1995). A realistic assessment of homology modeling efforts requires a more difficult test for side-chain prediction methods: placement of side-chains onto a related polypeptide backbone that is distinct from the native structure. This test has been performed less frequently and is usually applied to fewer than a dozen structures (Holm & Sander, 1992; Dunbrack & Karplus, 1993; Eisenmenger *et al.*, 1993; Wilson *et al.*, 1993; Koehl & Delarue, 1994). In contrast, we test our method on thousands of cases for the standard self-backbone test, but more importantly, we also characterize the performance of SCWRL on a large number of homology modeling test cases using template backbones both with and without gaps in the sequence alignment. SCWRL was tested and refined on a test set of 299 high resolution crystal structures, and then used to model every protein structure in the database of known structures. We then used SCWRL to create 9424 homology models, from available pairs of structures with similar sequences to test its performance as a homology modeling tool.

Results and Discussion

Tests on self-backbones

The SCWRL algorithm is described in detail in Methods. In brief, side-chains are placed on an experimentally derived or modeled protein backbone using the most probable rotamers from a backbone-dependent rotamer library. For this purpose, the 1996 update of the backbone-dependent rotamer library of Dunbrack & Karplus (1993) was used. Steric clashes are relieved systematically by a combinatorial search of rotamers in an order defined by the rotamer library and a modified van der Waals interaction energy.

We tested the SCWRL algorithm on a set of 299 crystal structures with resolutions better than or equal to 2.0 Å, *R* factors below 20%, sizes between 40 and 300 residues, and pairwise sequence identities less than 90%. For the 299 structures in the test set, the mean solution time was 37.6 seconds per structure, or 0.28 seconds per residue on a Silicon Graphics R4400 150 MHz processor.

Predictions for side-chain dihedral angles have traditionally been considered correct when they are within 40° of the crystal structure angles (Summers *et al.*, 1987; Lee & Subbiah, 1991; Dunbrack & Karplus, 1993; Koehl & Delarue, 1994; Hwang & Liao, 1995; Vasquez, 1995). We have followed that convention here. Table 1 shows the results of applying SCWRL to the test set of 299 high resolution structures, using a nine rotamer χ_1 and χ_2 backbone-dependent library. The conditional probability of χ_2 predicted correctly given a correct χ_1 is also shown. SCWRL, like other methods based on packing (Holm & Sander, 1992; Koehl & Delarue, 1994; Laughton, 1994), accurately predicts the conformation of the aromatic residues Phe and Tyr (χ_1 correct >90%). SCWRL also predicts χ_1 accurately for the γ-branched residues Leu, His, and Trp, and the β-branched residues Val, Thr, and Ile (χ_1 correct >80%), in addition to Pro. These residues exhibit strong interdependence between backbone and side-chain conformations. SCWRL performs less well for Ser, and the long, unbranched residues Met, Glu, Gln, Arg, and Lys.

Table 1. SCWRL results for 299 high resolution structures

Residue type	$\%\chi_1$ correct	$\%\chi_{1+2}$ correct	$\%\chi_{(2\|1)}$	$\%\chi_1$ consistent with rotamer model	$\%\chi_{1+2}$ consistent with rotamer model	Number of residues
Asn	73.3	51.6	70.4	96.6	96.6	2522
Asp	75.9	61.3	80.8	96.9	96.9	2808
Arg	64.9	48.5	74.7	93.8	86.4	1875
Cys	74.0			99.0	99.0	1258
Gln	68.2	49.2	72.1	93.6	89.8	1768
Glu	62.9	43.7	69.5	93.6	89.3	2599
His	85.3	73.4	86.0	99.3	96.2	988
Ile	87.4	69.8	79.9	98.0	94.0	2414
Leu	83.0	70.2	84.6	94.1	91.2	3650
Lys	67.5	49.4	73.2	92.4	85.2	3164
Met	71.7	51.1	71.3	96.6	94.6	869
Phe	90.4	84.1	93.0	99.3	91.0	1803
Pro	87.0	73.0	83.9	99.9	99.7	2057
Ser	62.1			97.2	97.2	3448
Thr	83.2			98.3	98.3	3050
Trp	87.4	63.1	72.2	99.3	94.3	732
Tyr	90.2	85.7	95.0	99.5	93.7	1857
Val	82.8			97.6	97.6	3401
Total	76.8	65.6	85.4	96.6	93.9	40263

Percentage correct for the 299 protein high resolution self-backbone test set is given for dihedral angles predicted within 40° of the crystal structure values. The percentages of crystal structure side-chains in conformations within 40° of any rotamer in the backbone-dependent rotamer library, which are therefore predictable by our method, are also shown.

Backbone-dependent side-chain rotamer libraries capture the steric restrictions imposed upon χ_1 rotamers by particular backbone conformations. As side-chain placement algorithms all consider the steric consequences of a particular rotamer, it is possible that backbone-dependent libraries will offer little advantage over their backbone-independent counterparts. To examine this possibility, we modeled our test set structures with three different rotamer libraries, a backbone-independent rotamer set, a backbone-dependent set containing three rotamers per residue (χ_1), and a backbone-dependent set containing up to nine rotamers per residue (χ_1 and χ_2). The results are shown in Figure 1. Marked improvements in using a backbone-dependent library over a backbone-independent library are shown for Ser, Val, Thr, Ile, Asp, Asn, and Pro. The backbone-independent library performs nearly as well as the backbone-dependent libraries for Phe and Tyr. The resolution of steric conflicts appears to be sufficient for positioning these side-chains. Providing additional choices for χ_2 also improves the prediction of both χ_1 and χ_2 for almost all residues. Even Cys, Val, and Thr, which have only χ_1 angles, are improved when using the nine rotamer backbone-dependent library. Presumably, this is due to improved positioning of neighboring residues. Exceptions to this improvement are seen for Ser, which may be too small to be affected by steric exclusions, Pro, which is well positioned using the backbone, and Asp and Asn, which are both small and usually near the surface of the protein.

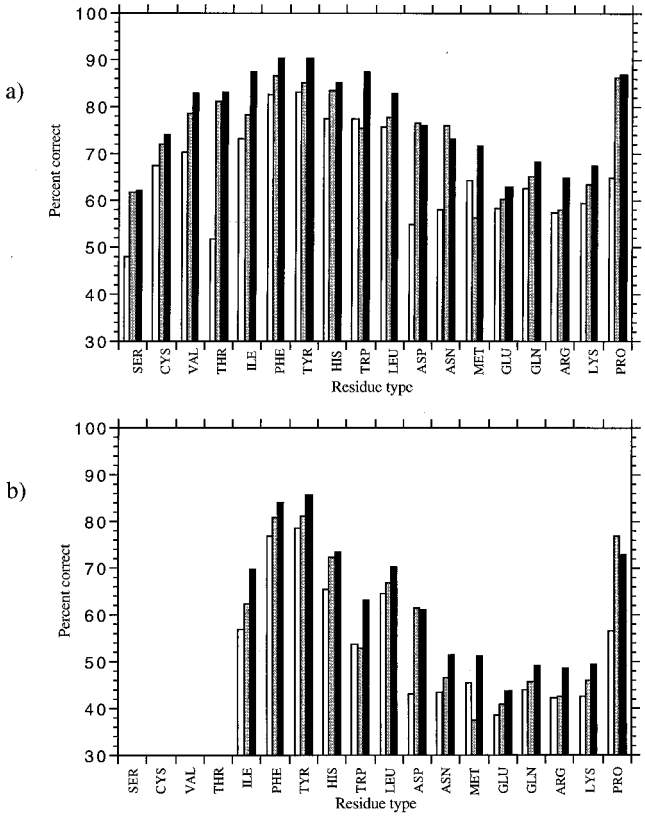

Figure 1. Comparison of different rotamer libraries for the high-resolution 299 protein self-backbone test set. Results for the backbone-independent library are shown in white, the three-rotamer backbone-dependent library in gray, and the nine-rotamer backbone-dependent library in black. Percent correct within 40° for χ_1 (a) and χ_{1+2} (b) are plotted for each residue type.

We were interested to see if our search method, which moves residues away from their most favorable backbone-dependent rotamers, significantly affected our results. In Figure 2, we compare the results for the three stages of the rotamer search. The residues Ser, Cys, Val, Thr, Asp, Asn, and Pro showed little or no improvement over the course of the search, indicating that they were either already well-positioned by considering their backbone conformation, or were too small to be affected by steric conflicts. The prediction accuracy for the other residues improved over the course of the rotamer search procedure. For the aromatic residues and Leu, the result was ~20% more accurate. These residues, which constitute a large portion of the core of a protein structure, seem to rely heavily on steric packing to determine their conformation.

Side-chain conformations are determined by a number of factors, including intrinsic conformational preferences, interactions with other parts of the protein, and interactions with solvent. We measured the level of solvent exposure of the residues in our test set and divided them into buried residues (<20% of their potential surface area exposed) and exposed residues. Because SCWRL does not consider interactions with solvent, and because exposed side-chains may exist in multiple conformations, we expect the predictions for exposed residues to be less accurate than those for buried residues. The results are shown in Figure 3. As expected, buried residues are predicted with much greater success than exposed residues, with significant improvements shown for buried Glu, Gln, Arg, and Lys residues.

How close are our predicted side-chain angles to the experimentally determined values? Although 40° is a traditional cutoff for side-chain prediction accuracy, we measured the absolute differences in χ_1 from the structures predicted by SCWRL and the crystal structures in the test set to evaluate the magnitude of the local distortions. The results are shown in Figure 4. For most side-chains, the predicted value of χ_1 is within ten degrees of the experimental value. These results also show that 40° is a suitable cutoff for measuring side-chain prediction, as this offers a clean separation between the correct rotamer bin, and the other choices which peak approximately 120° away. Figure 4 also illustrates that for many residues, there is a small but appreciable number of side-chain conformations recorded in crystal structures that are far from their rotamer values. These conformations represent ''non-predictable'' residues for our method, as they are more than 40° from any of their poten-

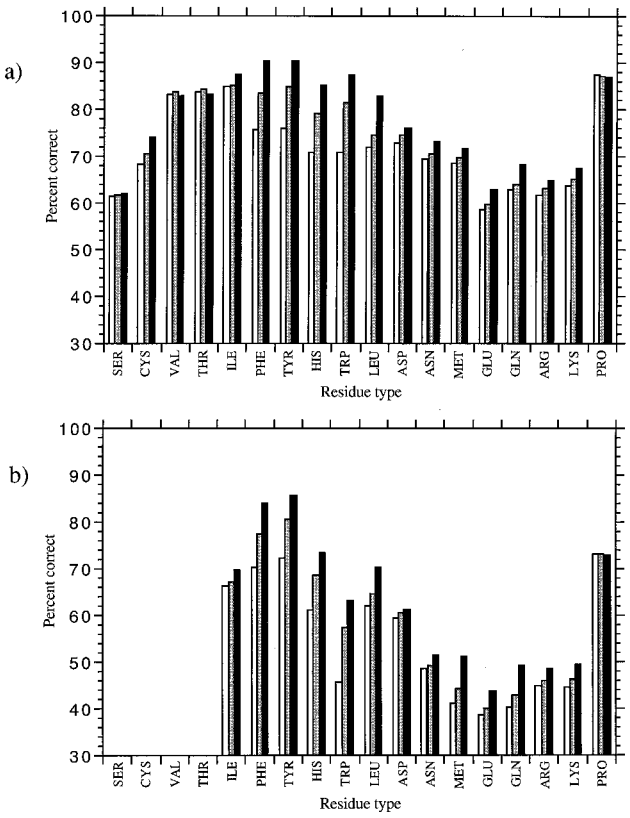

Figure 2. Comparison of results for the 299 protein self-backbone test set at different stages in the search process. Stage 1, in which the most favored backbone-dependent rotamer is built onto the backbone with no regard for steric conflicts, is shown in white. Stage 2, in which new rotamers are chosen to relieve side-chain/main-chain conflicts, is shown in gray. Stage 3 results, shown in black, are for the final structures after the relief of side-chain/side-chain conflicts by a combinatorial search. Percent correct within 40° for χ_1 (a) and χ_{1+2} (b) are plotted for each residue type.

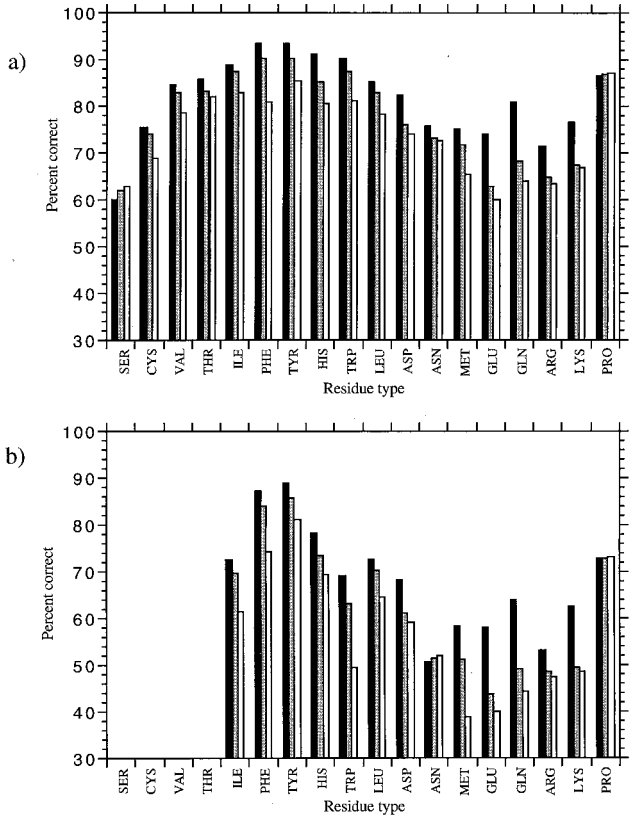

Figure 3. Effect of solvent exposure on prediction accuracy for the 299 protein self-backbone test set. Buried residues, those with less than 20% of their potential surface area exposed, are shown in black, exposed residues in white, and all residues in gray. χ_1 (a) and χ_{1+2} (b) are plotted for each residue type.

tial rotamer replacements. Table 1 shows the percentage of the residues in our test set population which are within 40° of any of the rotamers in the library, and which are therefore consistent with the rotamer model for side-chains and predictable by our method. Examples of conformations which are inconsistent with a rotamer model are particularly numerous for long polar side-chains. These conformations, with χ_1 within 20° of 0°, 120°, and −120°, for example, are unlikely to be correct, since the strain energy involved is greater than 4 kcal/mole (Durig & Compton, 1979; Compton *et al.*, 1980; Wiberg & Murcko, 1988). Some of these cases may be due to crystallographic averaging of two rotamers. For example, if a residue exists in a crystalline state with a population of roughly half with χ_1 ~60° and half with χ_1 ~180°, the observed χ_1 may be 120° (B. W. Matthews, personal communication).

In the course of modeling protein structures from the entire PDB, we encountered many crystal structures of modest resolution and structures determined by NMR spectroscopy in which some side-chain assignments are likely to be in error. This type of error was observed by Morris *et al.* (1992), who noted that several parameters indicate the decrease in stereochemical quality of crystal structures as resolution decreases. We were interested to find a correlation between the resolution of a crystal structure and the accuracy of our method. Table 2 shows our results for these structures, broken down by resolution. For NMR structures, the first structure in the PDB file was taken to be the reference structure. SCWRL side-chain predictions are least compatible with NMR structures. In addition, there is a clear correlation between the resolution of a crystal structure and SCWRL's ability to predict the rotamer positions for all side-chain types. We measured this effect more directly by identifying pairs of crystal structures of identical proteins with differences in resolution $\geqslant$0.5 Å. We used the higher resolution structure as the correct reference structure for side-chain positions. For the 41 pairs of structures, in which the average difference in resolution was 0.76 Å, χ_1 accuracy improved an average of 5.4% for the high resolution over the low resolution backbone. For the eight

Table 2. Effect of resolution on prediction accuracy for the database of known structures

Residue type	Resolution 0.0 to 2.0 (Å)	2.0 to 2.5 (Å)	2.5 to 3.0 (Å)	3.0 and up (Å)	NMR	All structures	No. residues
Asn	70.8	69.4	64.7	58.4	42.1	66.7	59554
Asp	77.9	74.3	67.7	59.7	44.1	71.2	72679
Arg	66.6	64.1	59.3	52.0	49.1	61.7	56414
Cys	75.8	72.6	67.1	64.0	49.2	69.3	22240
Gln	69.1	66.1	61.2	56.3	49.2	63.9	45962
Glu	65.3	59.9	55.1	50.6	48.6	58.4	72385
His	84.0	82.7	78.2	72.8	54.9	80.0	28955
Ile	87.7	85.4	80.9	73.6	66.3	83.1	66503
Leu	84.4	79.5	74.8	69.0	62.3	77.8	104678
Lys	69.6	65.5	60.9	55.7	50.0	63.8	77316
Met	70.6	69.5	64.8	59.0	45.8	66.6	24442
Phe	91.3	89.8	87.3	83.8	71.5	88.3	49481
Pro	84.7	81.0	76.6	76.2	88.3	80.4	57098
Ser	63.6	57.3	49.6	44.9	33.7	54.8	83024
Thr	83.0	77.3	69.3	60.5	48.7	73.9	79397
Trp	90.1	86.3	83.6	74.8	59.7	84.8	19317
Tyr	89.9	89.2	85.6	80.9	65.3	86.9	45474
Val	83.8	79.1	72.9	66.2	63.4	76.9	87872
χ_1 total	77.7	74.1	68.7	63.3	54.3	71.8	
χ_{1+2} total	65.9	61.1	54.9	49.4	38.8	58.7	
No. structures	1296	1345	709	245	590	4185	
No. residues	243464	401292	271289	98572	38074	1052791	

pairs with differences in resolution $\geqslant$1.0 Å, the average improvement in χ_1 accuracy was 13.3%. In the cases of low resolution X-ray structures, it is possible that SCWRL is predicting the correct rotamers for some positions that are incorrectly reported in the experimental structures. In light of this, SCWRL may be a useful tool for building side-chains onto a main-chain trace in the course of experimental structure refinement.

Based on the observation that crystal structures of identical proteins can contain residues in different rotamer positions (Faber & Matthews, 1990; Kossiakoff *et al.*, 1992; Kishan *et al.*, 1994), we attempted to measure the amount of variability for different residue types in proteins that had been solved in different crystal structures. This would put an upper limit on the accuracy that any side-chain prediction program could hope to achieve. We began by searching for sets of proteins with identical sequences solved in different crystal forms in structures with resolution better than or equal to 2.0 Å. Additional structures were included as long as they were not solved by the same investigators, in order to reduce the chance of including structures that were refined using another crystal structure for molecular replacement. Our final set of structures for this study included: hen egg white lysozyme (PDB references 1VFB, 132L, 2LZT, 1HEW, 1LSE, 1HEL, 1LZA, 2LYM, 6LYT, 1LMA, 1LYS, 4LYT, and 5LYM) including structures in five different space groups; bacteriophage T4 lysozyme (177L, 178L, and 179L) in two different space groups; and bovine pancreatic trypsin inhibitor (4PTI, 5PTI, 6PTI, and 1BPI) in two different space groups. Each residue position in each of these sets was measured to determine whether the side-chains remained in the same χ_1 and χ_{1+2} rotamer bins, or whether these positions varied. For example, of the 18 Ile residues in the three sets of structures, 14 displayed the same χ_1 bin values in every structure, and ten displayed the same χ_{1+2} rotamers. The results are shown in Figure 5, along with the prediction results of SCWRL on the test set. The variability in the crystal structure rotamers are shown as values with 95% confidence intervals, derived graphically from the binomial distribution for small sample sizes (Clopper & Pearson, 1934; Glantz, 1992). Cys residues involved in disulfide bonds in these structures were omitted from the analysis and no free Cys residues were found. As

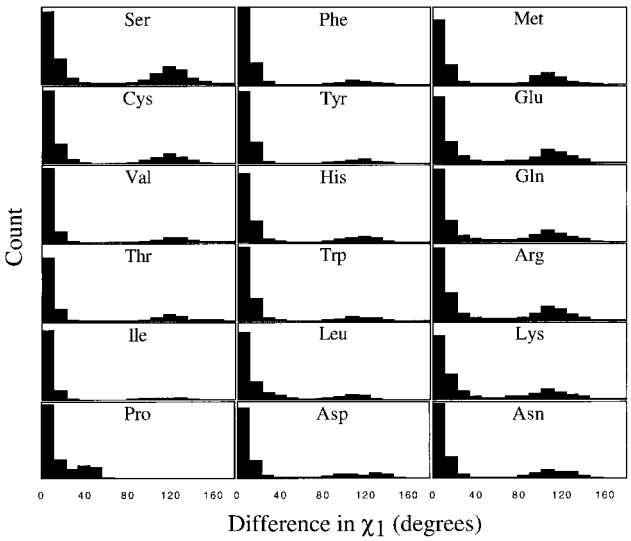

Figure 4. Absolute differences in the χ_1 values predicted by SCWRL and the crystal structure values for the 299 protein self-backbone test set.

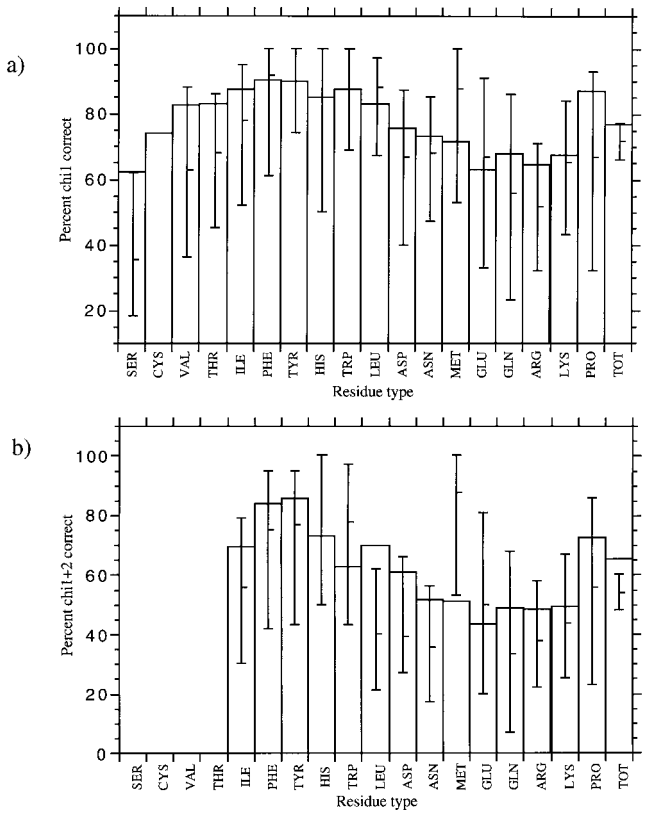

Figure 5. The results for the self-backbone test set of 299 structures (bars) are shown with the 95% confidence intervals for a maximum limit of prediction accuracy, based on multiple crystal structures of identical proteins, for χ_1(a), and χ_{1+2}(b).

more structures are solved in multiple crystal forms, our confidence in these limits will improve. However, the initial results in Figure 5 indicate that SCWRL is working up to its theoretical limits for most residue types, even without consideration of hydrogen bonding, solvent effects, and electrostatic interactions.

In an effort to improve predictions for the more problematic polar residues, we tested a hydrogen bonding term for Ser, Asp, Asn, Met, Arg, Lys, Glu, and Gln. For these residues, rotamers which could potentially make a hydrogen bond with the protein backbone up to six residues away from the side-chain were given a higher rank than rotamers which could not make such a hydrogen bond. Including this term did not have significant effect on the results for these residues, with changes in accuracy for χ_1 prediction of less than 1%, except for Ser which decreased by 4%. Other variations of our method for predicting these side-chains were tested, including increasing the radii of the side-chain atoms, and truncating Met, Arg, Lys, Glu, and Gln residues at C^{γ} or C^{δ}. These changes also did not improve our results (data not shown).

A limited energy minimization refinement is often used in homology modeling after the placement of side-chains. We performed energy minimizations with the program CHARMM (Brooks *et al.*, 1983) on our test set structures to see if a short, constrained minimization would change our predictive ability. Minimizations were performed with atoms constrained with harmonic potentials to their Cartesian coordinates. The constraints on the side-chain atoms were reduced every 50 steps of adopted-basis Newton Raphson minimization; the final 50 steps were performed with no constraints applied to the side-chain atoms (Summers & Karplus, 1989; Dunbrack & Karplus, 1993). We found that fewer than 2% of SCWRL-placed side-chains changed either χ_1 or χ_2 by 40° or more, leaving our scores practically unaltered (data not shown). This may be due to our use of rotamers close to local energy minima.

Comparison with other methods

Many algorithms have been developed to predict side-chain conformation. To be useful, a new approach must be more accurate and/or more efficient than existing methods over a large set of test cases. Figure 6 compares the results of SCWRL with those from the Monte Carlo program torso from the MAXSPROUT package of Holm & Sander (1991), and with the results from the mean field theory algorithm described by Koehl & Delarue (1994), on 262 structures from the test set. The remaining 37 structures from the test set contained incomplete backbones, which are handled differently by the three methods and are therefore not readily comparable. We chose to compare SCWRL with these algorithms because they are both fast, publicly available, and accurate. We attempted to make this comparison as consistent as possible by using the same test structures and evaluation criteria for all three methods. Both the Holm & Sander and Koehl & Delarue algorithms use the backbone-independent library of Tuffery *et al.* (1991). The accuracy of SCWRL and torso are comparable. SCWRL performs slightly better for Pro, χ_1 of Asp, and χ_2 of Leu, which displays a strong dependence on χ_1 (Dunbrack & Karplus, 1994). In contrast, torso does much better for Met at both χ_1 and χ_2. SCWRL displays a distinct advantage over the Koehl & Delarue algorithm for Ser, Thr, Pro, the χ_1 of Asp, and the χ_2 of Leu. The Koehl & Delarue algorithm, again using the Tuffery library, shows an advantage for Met residues. The final values for χ_1 and χ_{1+2} predicted correctly for all residues are 72.6% and 60.9% for the Koehl & Delarue algorithm, 76.7% and 64.3% for the Holm & Sander algorithm, and 77.0% and 65.8% for SCWRL.

Many authors have used root mean squared deviation of side-chain atom positions as an evaluation for their side-chain placement methods (Lee & Subbiah, 1991; Tuffery *et al.*, 1991; Holm & Sander, 1992; Levitt, 1992; Dunbrack & Karplus, 1993; Eisenmenger *et al.*, 1993; Wilson *et al.*, 1993; Koehl & Delarue, 1994; Laughton, 1994; Hwang &

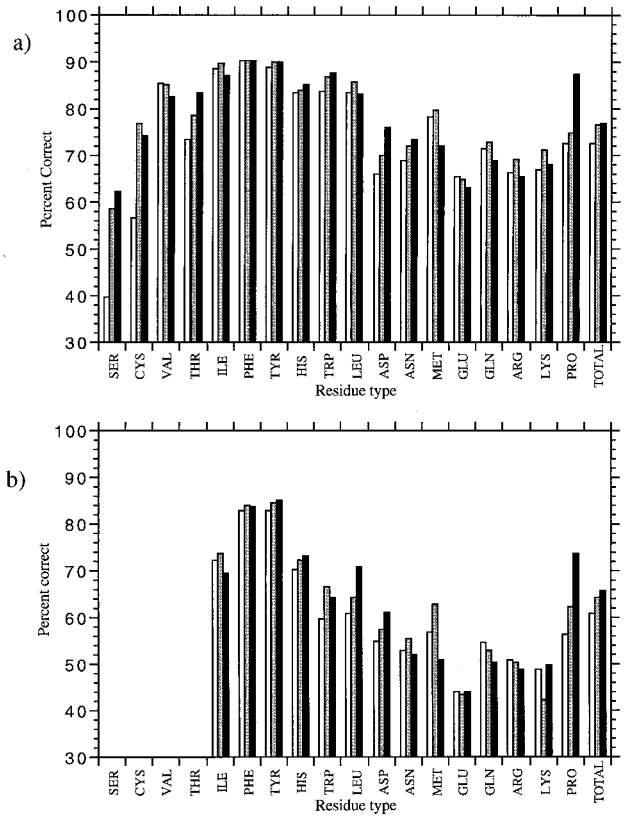

Figure 6. Comparison of SCWRL results in black with the Monte Carlo sampling algorithm of Holm & Sander (1991) in gray and the mean-field algorithm of Koehl & Delarue (1994) in white for 262 chains of the 299 protein self-backbone test set. Percent correct within 40° for χ_1 (a) and χ_{1+2} (b) are plotted for each residue type.

Liao, 1995; Vasquez, 1995; Chung & Subbiah, 1996). We present the results for this evaluation in Table 3, with the results for the Koehl & Delarue and Holm & Sander algorithms. These scores include all side-chain atoms beyond C^{β}, which makes them larger than scores which include C^{β} (Shenkin *et al.*, 1996), and we correct for crystallographically symmetrical side-chains (Asp, Asn, Phe, Tyr, His). The total root mean squared deviation score per structure, which has also been used historically, is also presented in Table 3.

We would expect methods that use a backbone-independent rotamer set, like the Koehl & Delarue and Holm & Sander methods, to perform less well on exposed residues than methods that use backbone-dependent rotamer sets, like SCWRL. Obviously, a side-chain which is exposed to solvent is not subject to the steric exclusion of as many other side-chain and main-chain atoms from the protein. In these cases, the local backbone conformation may become the dominant factor influencing a side-chain's conformation. When the results from the Koehl & Delarue and Holm & Sander methods are divided by solvent exposure of the residues and compared with the results from SCWRL, we see that this expectation is borne out. Figure 7 shows these results for correct χ_1 predictions. While the Holm & Sander algorithm shows an advantage for buried residues, SCWRL performs better for exposed residues. Because exposed residues comprise sites on a protein which interact with other molecules, including small-molecule binding sites, protein-protein recognition sites, and enzymatic active sites, the ability to predict these side-

Table 3. r.m.s. deviations of side-chain atoms on 262 high resolution structures

Residue type	Koehl & Delarue rmsd (Å)	Holm & Sander rmsd (Å)	SCWRL rmsd (Å)	Number of residues
Asn	1.345	1.253	1.281	2280
Asp	1.370	1.255	1.106	2475
Arg	2.729	2.760	2.835	1688
Cys	1.427	0.888	0.990	1140
Gln	1.603	1.677	1.707	1612
Glu	2.136	2.119	2.161	2314
His	1.235	1.202	1.154	884
Ile	0.707	0.697	0.804	2166
Leu	1.134	1.039	0.960	3272
Lys	2.033	2.319	2.065	2786
Met	1.697	1.555	2.054	782
Phe	1.043	0.980	1.009	1574
Pro	0.435	0.485	0.504	1829
Ser	1.471	1.079	1.018	3127
Thr	0.853	0.680	0.618	2725
Trp	2.110	1.751	1.916	645
Tyr	1.207	1.123	1.150	1677
Val	0.508	0.537	0.631	3072
Average rmsd per residue	1.327	1.254	1.250	36048
Average total rmsd per structure	2.007	1.964	1.931	262 structures

Root mean squared deviations of side-chain atoms for SCWRL, the Koehl & Delarue (1994) method, and the Holm & Sander (1991) method on a subset of the 299 protein self-backbone test set. These measurements include all side-chain atoms beyond C^{β}, and are corrected for crystallographically symmetrical residues (Asp, Asn, Glu, Gln, Phe, Tyr, His, Arg).

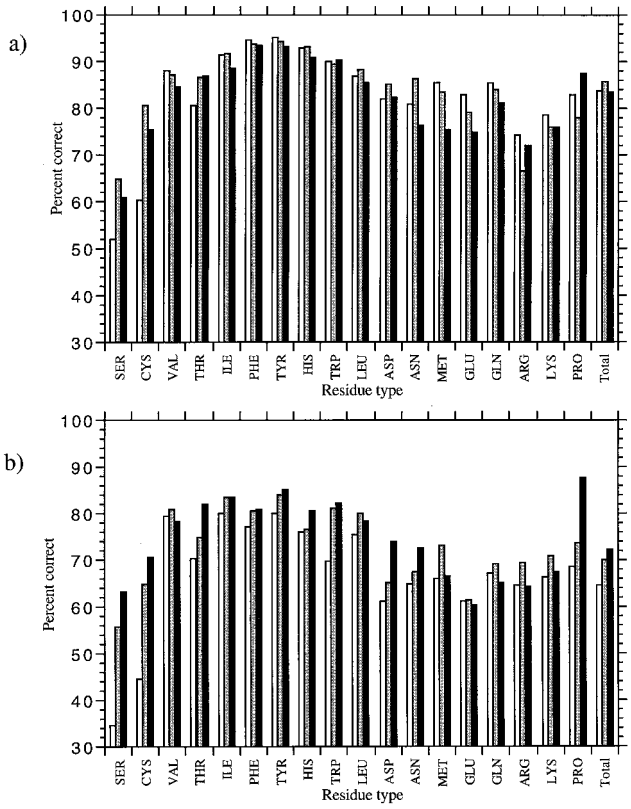

Figure 7. Comparison of the buried (a) and exposed (b) χ_1 results for the Holm & Sander (1991) algorithm in gray, the Koehl & Delarue (1994) algorithm in white, with SCWRL, in black for the 299 protein self-backbone test set. Buried residues are defined as having less than 20% of their potential surface area exposed.

chain conformations more accurately is important.

The curious peak in the predictive ability of both the Monte Carlo and mean field methods for Met led us to study the behavior of that residue with different rotamer sets. We first tested the ten backbone-independent rotamers for Met from the Tuffery *et al.* (1991) library in SCWRL. Then, we tested a full backbone-dependent library for Met which articulated χ_3 as well as χ_1 and χ_2, with up to 17 rotamers per position. Neither of these rotamer sets improved our accuracy for Met (data not shown). Some confluence of the backbone-independent rotamer set and the search algorithms for these methods must create this increased accuracy.

Homology modeling: prediction of side-chain conformations on non-self backbones

A level of sequence identity above 25% between proteins often indicates a structural similarity which can be exploited in homology modeling (Blundell *et al.*, 1987). This structural similarity imparts a finite amount of side-chain information to a homology modeling effort. We tested SCWRL by constructing homology models of two types: a set without insertions or deletions in the sequence alignments of the target and template proteins and a set allowing such gaps. Table 4 shows the percent χ_1 accuracy for a set of 1822 homology models without gaps built by SCWRL, first with no knowledge of template side-chain positions (method I), with conserved residues retaining their Cartesian coordinates (method II), and with χ_1 rotamers preserved from the template (method III) (Summers & Karplus, 1989; Dunbrack & Karplus, 1993). No pairs of sequences without gaps were found for crystal structures with percent identity between 30 and 40%. Adding the information for conserved side-chains in methods II and III improved the results by 5 to 10% over method I. The most striking result is that the predictive ability of SCWRL using homology modeling method II was better than using SCWRL on the high resolution test set, even though the latter case uses the native backbone, while the former relies upon a surrogate backbone scaffold. The addition of preserved χ_1 bins from the template in method III improved the results as much as method II. A fourth method, which combined the exact conformers of conserved residues from method II and the preserved χ_1 bins for mutated residues from method III, was tested but showed no improvement over methods II and III (data not shown). The final values for χ_1 and χ_{1+2} correct in the high resolution, no gaps test cases are 73.5% and 60.2% for method I, 81.8% and 71.8% for method II, and 81.9% and 66.6% for method III, respectively. Because the template side-chain information is usually available during a real homology modeling effort, the accuracy of SCWRL supports its utility as an homology modeling tool.

Because the backbone resolution affected the prediction accuracy for the self-backbone models from the PDB (see Table 2), we measured the effect of both the template resolution and the target resolution on the prediction accuracy for the first set of homology models (no gaps). Table 4 shows how the prediction accuracy varies for homology models where both the target and template structures have resolution <2.0 Å, and homology models where one or both of the template and target structures have resolution >2.0 Å. Increased resolution of both target and template structures leads to increased prediction accuracy. A higher resolution template structure provides a more accurate backbone for the side-chain placement, while higher resolution target structures provide a more accurate description of the residues which are to be modeled.

We constructed another set of homology models from sequence pairs which aligned with gaps, a situation common in homology modeling. The sequences from the 299 test set structures were aligned, and those 267 pairs with greater than 30% sequence identity were used to create 534 homology modeling tests. Insertions in the target sequence were deleted, and insertions in the template structure were removed to create a matching template structure and target sequence.

Table 4. Homology modeling results

		Target and template resolution (Å)	Sequence Identity 30–40%	40–50%	50–60%	60–70%	70–80%	80–90%	90–100%	Total
Method I	No gaps	>2.0		62.1	69.1	69.1	70.6	68.4		68.4
	No gaps	<=2.0		70.3	77.1	74.8	75.7	72.0		73.5
	Gaps allowed	<=2.0	61.2	63.0	60.2	68.5	70.8	70.3	72.8	64.7
Method II	No gaps	>2.0		67.8	74.3	76.6	75.9	76.7		75.7
	No gaps	<=2.0		70.0	83.1	82.5	82.8	81.4		81.8
	Gaps allowed	<=2.0	64.5	66.7	65.0	74.9	79.8	81.2	81.0	69.8
Method III	No gaps	>2.0		65.7	73.1	75.8	77.1	77.4		75.8
	No gaps	<=2.0		72.7	82.4	81.9	84.2	82.1		81.9
	Gaps allowed	<=2.0	63.0	65.3	64.6	74.6	79.7	81.1	81.3	68.9
Number of models	No gaps	>2.0		124	134	370	170	522		1320
	No gaps	<=2.0		12	66	186	2	236		502
	Gaps allowed	<=2.0	142	152	44	82	28	34	52	534

Results for χ_1 accuracy are shown for three homology-modeling (non-self backbone) side-chain placement methods, as well as the number of models built with each method. Method I uses only the template backbone and target sequence to build side-chains, while methods II and III utilize side-chain information from the template structure to guide side-chain placement.

Each of the 534 tests were modeled using the three methods previously described. The results from this study are shown in Table 4, with percent identities given as the number of conserved residues over the aligned regions. Using alignments with gaps generally lowers the prediction accuracy approximately 10%, but the relative effectiveness of the different methods is retained.

Other methods for side-chain placement have not typically been tested on homology models (Tuffery *et al.*, 1991; Levitt, 1992; Laughton, 1994; Hwang & Liao, 1995; Vasquez, 1995), or have been tested on a dozen or fewer models (Holm & Sander, 1992 (ten models); Dunbrack & Karplus, 1993 (one model); Eisenmenger *et al.*, 1993 (two models); Wilson *et al.*, 1993 (four models); Koehl & Delarue, 1994 (12 models)). The accuracies reported for some of these models are similar to the accuracy of our results, but the statistical power of our large test sets gives us confidence in the reproducibility of our results.

Conclusions

We have developed a method for placing side-chains onto a protein backbone which is fast, easy to use, and more accurate than other fully automated, publicly available methods. The prediction accuracy for this method in the test case of building side-chains onto a native backbone approaches observed limits of accuracy for the underlying experimental data. The method also achieves a useful prediction accuracy in a test of almost 10,000 homology models. These results indicate that SCWRL is a potentially useful tool for real homology modeling projects.

SCWRL is publicly available to academic groups on the World Wide Web at http://www.cmpharm.ucsf.edu/cohen/ and by anonymous ftp at ftp.cmpharm.ucsf.edu. The backbone-dependent rotamer library and other information on conformational analysis of protein side-chains is available at http://www.cmpharm.ucsf.edu/ ~ dunbrack.

Methods

General

Protein structures were taken from the August 1996 update of the Brookhaven Protein Databank (Bernstein *et al.*, 1977). Computation was performed on a Silicon Graphics R4400 150 MHz processor. The ALIGN program from the FASTA package (Pearson & Lipman, 1988; Pearson, 1990) was used for sequence alignment. Graphical display and measurement of structures were performed using PSSHOW v.2.0d (Swanson, 1994).

We begin with the main-chain atoms N, C^{α}, C, and O from a protein structure. For self-backbone tests, we do not take any side-chain atoms as input. The sequence, together with the ϕ and ψ angles of the backbone, are used to look up an ordered list of rotamers for each residue from a rotamer database. Each of these potential rotamers is built, using bond lengths and angles from the AMBER 4.1 parameter set (Pearlman *et al.*, 1995), and the set of rotamers is searched for the minimum steric clash to create the output structure.

The energy function in SCWRL is a simple repulsive steric clash check. For a pair of atoms, the energy of interaction is given by:

Table 5. Radii used for steric clash checks

Atom type	Radius (Å)
Backbone N	1.1
Backbone C-alpha	1.3
Backbone C	1.3
Backbone O	1.1
Asp, Asn OD1, OD2, ND2	0.7
Glu, Gln OE1, OE2, NE2	0.7
Lys CE	0.8
Lys NZ	0.4
Met CE	0.8
Arg NE	0.7
Arg CZ	0.8
Arg NH1, NH2	0.4
Other side-chain C and S	1.3
Other side-chain O and N	1.1

$$E = 0.0 \qquad R > R_0$$

$$E = \left(\frac{R}{R_0}\right)(-57.273) + 57.273 \qquad R_0 \geq R \geq 0.83R_0 \qquad (1)$$

$$E = 10.0 \qquad R < 0.83R_0$$

where R is the distance between the atoms and R_0 is the sum of their radii. The linear portion of the function approximates the repulsive curve of a Lennard-Jones potential (see Figure 8). To make the steric term more forgiving on the rigid rotamers, the radii of atoms are reduced roughly 15% from their van der Waals radii (Schulz & Schirmer, 1979; see Table 5), to values which approximate the distance where a Lennard-Jones potential would become repulsive. In addition, the radii of terminal atoms of longer residues, whose positions are determined by χ_3 and χ_4, are further reduced. This is designed to lessen the impact of having fixed χ_3 and χ_4 values in our rotamer set.

Search strategy

To overcome the combinatorial nature of the side-chain packing problem, our search strategy does not involve a search of every rotamer of every side-chain, but rather takes a structure with residues in their most favorable backbone-dependent rotamers and systematically resolves the conflicts that arise from that structure. Each residue begins in its most favored rotamer, according to the rotamer database used. This is the first stage structure. When a side-chain from the first stage structure has a steric clash as defined by equation (1) with the (fixed) main-chain, the rotamer for that residue is changed to progressively less favorable rotamers until one is found that does not conflict with the main-chain. The second stage structure has all of these side-chain to main-chain clashes relieved (Summers & Karplus, 1989; Dunbrack & Karplus, 1993).

At this point, there will be some side-chain to side-chain clashes. Where these occur, the two clashing residues are put in a ''cluster'' of residues which interact. The clashing residues are allowed to explore all of their respective rotamers, save those which cause side-chain to main-chain clashes, and any residues which clash with these ''rotating'' residues are added to the cluster and themselves allowed to explore all their rotamers. In this way, clusters of residues grow by clashing with residues not in the original cluster. Clusters can also merge when two rotamers from residues in different clusters clash, in which case all of the residues from the two clusters are combined to form a single larger cluster. Side-chains which do not clash with the main-chain, and are never involved in a steric clash with an activated residue, are left in their most favorable backbone-dependent rotamers.

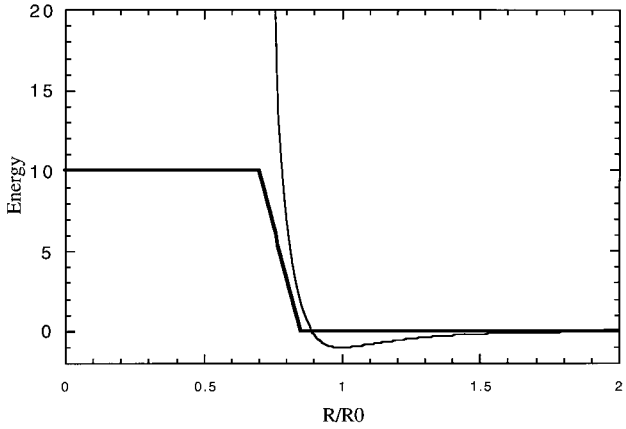

Figure 8. Plot of the energy function for steric clash checks in SCWRL (dark line), with a standard Lennard-Jones potential (light line). The energy function is shown relative to ''full size'' atomic radii. SCWRL atomic radii are reduced approximately 15%.

An example of the stepwise solution to side-chain backbone and side-chain/side-chain steric clashes is illustrated in Figure 9 for a 50-residue mouse protein kinase C, PDB entry 1PTQ (Zhang *et al.*, 1995).

When the clusters have grown to their final size, each one represents an exclusive subset of residues which are allowed to interact with each other. Each cluster is solved, in turn, through a combinatorial search to find its minimum steric clash score. When all of the clusters have been solved, the stage three structure is output as the solution.

The search procedure tests each residue and combination of residues in the order they were added to the cluster. Rotamers for each residue are tested in order of decreasing favorability. The first combination of rotamers with a steric clash score of zero is taken as the final solution. If no such combination is found, all the rotamer combinations are searched, and the combination with the minimum steric clash score is taken as the best possible solution.

Parsing large clusters

Occasionally, clusters grow too large to be solved quickly with a combinatorial search. When such a large number of combinations is reached, the cluster is broken into sub-clusters to speed the solution time. In our case, the limit is set for clusters that cannot be solved by the combinatorial search in approximately one second, which is reached for clusters containing more than 1.5×10^7 rotamer combinations, about 15 residues. A large cluster is parsed by finding the residue in that cluster whose removal from the cluster results in the smallest sub-clusters. Then each of the sub-clusters is solved in the presence of each of the ''keystone'' residue's potential rotamers. For example, in a 21-residue cluster where every residue has three potential rotamers, the combinations to search will number 3^{21}, or 1.0×10^{10}. If a residue in this cluster is found which divides the cluster into

Stage 1: All side-chains placed according to the rotamer library

Stage 2: Side-chains for Y6 and L29 conflict with the backbone. The preferred order of rotamers for the local backbone of Y6 is g^-, g^+ and t. Because the g^- rotamer conflicts with the backbone, the g^+ rotamer is tested and also found to conflict with the backbone. The t rotamer is then tested, and does not conflict with the backbone. Y6 is placed in the t rotamer.
The ordered rotamers for L29 are tg^+, g^-t, tt, and g^-g^+. The tg^+ rotamer conflicts with the backbone, so the g^-t rotamer is tested and found not to conflict. L29 is placed in the g^-t rotamer.

Stage 3: Initially, four side-chains are involved in side-chain to side-chain conflicts:
K4 conflicts with E32
Y6 conflicts with K30
These form the basis for two clusters of interacting residues. All rotamers for these four residues are checked, and residues Y8, L21, and N37 are added to the second cluster due to steric conflicts with rotamers of Y6 and K30.
The two clusters are solved in turn: cluster one is solved by

K4 g^-t to g^-g^-
E32 g^-t to g^-t

The second cluster is solved by

Y6 t to t
K30 g^-t to tt
Y8 g^- to g^-
L21 g^-t to g^-t
N37 g^-g^- to g^-g^-

Figure 9. Outline for the resolution of steric conflicts for PDB entry 1PTQ (Zhang *et al.*, 1995).

two 10-residue sub-clusters, then the combinations to search will number $3(3^{10} + 3^{10}) = 3.5 \times 10^5$.

This parsing of clusters into sub-clusters is a recursive process, and if the sub-clusters still contain more combinations than the cutoff, or if a keystone residue fails to break a cluster into non-interacting sub-clusters, the remaining clusters are passed down to a new level for additional parsing. In some very rare cases, the parse routine cannot find a subset of keystones which breaks the cluster up fast enough to overcome the combinatorics.

Rotamer libraries

Three different rotamer libraries were tested. To measure the importance of the backbone-dependence of the rotamers, the first library is a set of backbone-independent rotamers derived from the 393 protein chains used in the 1996 update of the rotamer library described by Dunbrack & Karplus (1993). This set represents each side-chain in each of three positions , t, g^+, and g^-, at χ_1 for all side-chains except Pro and at χ_2 for Arg, Lys, Met, Glu, Gln, Asp, Asn, Trp, Ile, and Leu, producing a total of nine rotamers for these side-chains. Pro χ_2 is determined by its χ_1 rotamer, while the aromatics have χ_2 set at their most favorable position, near 90°. For the longer residues Met, Glu, Gln, Arg, and Lys, χ_3 and χ_4 are left in the *trans* rotamer. The rotamers for each residue type are listed with their observed frequency in the database. This set represents a backbone-independent rotamer library akin to that described by Ponder & Richards (1987).

The second set of rotamers is a backbone-dependent library which comes from the 1996 update of the Dunbrack and Karplus rotamer library. This library provides three rotamer choices for each residue, with χ_1 in t, g^+, or g^- positions. The rotamers are listed with their observed frequencies, or with their expected frequencies in backbone positions that are sparsely populated in the database. Frequencies were determined with a Bayesian statistical analysis of the backbone-dependent rotamer populations (R. L. D. & F. E. C., unpublished results). χ_2 for this library is typically left in the *trans* rotamer, but the exact value varies with dependencies on χ_1 and the local backbone. χ_3 and χ_4 for longer side-chains are set to 180°.

To test the advantage of articulating χ_2 as well as χ_1, a third library was constructed. This library also contains backbone-dependent rotamers, but with up to nine rotamers representing t, g^+, and g^- positions for χ_1 and χ_2. The rotamers were ordered based on observed and expected frequencies from the Bayesian analysis of the backbone-dependent χ_1 rotamer frequencies and the con-

ditional probabilities of the χ_2 rotamers on the χ_1 rotamers.

Self-backbone tests of side-chain conformation prediction

We evaluated SCWRL by predicting side-chain conformations from the experimentally determined backbones of 299 crystal structures. These were protein crystal structures from the PDB with resolutions <2.0 Å and *R*-factors <20%, ranging from 40 to 300 residues, with sequence identities less than 90% between any two structures in the set.

In a second evaluation stage, SCWRL was applied to all 4705 structures in the August 1996 update of the PDB. Of these, 338 contained solely nucleic acid or HETATOM records and 124 were theoretical model structures. Of the 4243 which remained, 58 failed to pass through SCWRL successfully, either because they contained more than 2000 residues (24) or because the parse subroutine failed to break the clusters into manageable size (34). The remaining 4185 structures were successfully modeled by SCWRL.

Prediction of side-chain conformations in homology (non-native backbone) modeling

Pairs of protein sequences from every protein structure in the PDB were aligned. The 911 pairs of alignments from sequences of the same length, which aligned with no gaps and which had 30 to 90% sequence identity, were used by SCWRL to construct 1822 homology models with the sequence of one from each pair modeled on the backbone of the other, and *vice versa*. The first 1822 models were constructed using only the backbone coordinates from the template structure. This set is labeled method I. These models were then recalculated, retaining the Cartesian coordinates of residues from the template crystal structure where they were identical in the sequence alignment to the target structure, for example where a Phe from the target sequence matched a Phe from the template sequence. All mutated residues, for example a Phe matched against an Arg, were constructed from the backbone-dependent χ_1 and χ_2 library. This second set of 1822 models is labeled method II. A third set of models were built, retaining the χ_1 rotamers between substitutions of residues containing the same number of atoms in the C^γ positions. For example, a template Arg determines the χ_1 rotamer of the target His, and a template Thr determines the χ_1 rotamer of the modeled Ile in the same position. For substitutions from template Pro, Gly, or Ala residues to other residues, the library is used. This set is labeled method III. A method IV set was constructed with the rules of method III, except that identical residue side-chain coordinates are preserved as in method II.

A second set of 534 homology model targets, this time with gaps, was created by aligning the 299 test set sequences and retaining those pairs with greater than 30% sequence identity. The test set was used to take advantage of the high quality of the structures, as well as to limit the number of homologous pairs found, which would be enormous in the full PDB when gaps are allowed. Deleted residues, where the template is longer than the target, were removed from the template structure while insertions, where the target is longer than the template, were not modeled. The four modeling methods described above were used on each of these models in this test set.

Acknowledgements

We thank John Troyer for helpful programming advice. M. J. B. is partially supported by an NIH pharmaceutical sciences training grant (GM07175). R. L. D. is an NIH postdoctoral fellow (GM16279). This work was supported by a grant from the National Institutes of Health (GM39900).

References

Bamborough, P. & Cohen, F. E. (1996). Modeling protein-ligand complexes. *Curr. Opin. Struct. Biol.* **6**, 236–241.

Benedetti, E., Morelli, G., Nemethy, G. & Scheraga, H. A. (1983). Statistical and energetic analysis of side-chain conformations in oligopeptides. *Int. J. Peptide Protein Res.* **22**, 1–15.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bhat, T. N., Sasisekharan, V. & Vijayan, M. (1979). An analysis of side-chain conformation in proteins. *Int. J. Peptide Prot. Res.* **13**, 170–184.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature,* **326**, 347–352.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

Browne, W. J., North, A. C. & Phillips, D. C. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* **42**, 65–86.

Chandrasekaran, R. & Ramachandran, G. N. (1970). Studies on the conformation of amino acids, XI. Analysis of the observed side group conformations in proteins. *Int. J. Protein Res.* **2**, 223–233.

Chung, S. Y. & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure,* **4**, 1123–1127.

Clopper, C. J. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika,* **26**, 404.

Compton, D. A. C., Montero, S. & Murphy, W. F. (1980). Low-frequency Raman spectrum and asymmetric potential function for internal rotation of gaseous n-butane. *J. Phys. Chem.* **84**, 3587–3591.

David, C. (1993). Sprouting side-chain conformations in X-PLOR simulations of peptides. *J. Comput. Chem,* **14**, 715–717.

Defay, T. & Cohen, F. E. (1995). Evaluation of current techniques for ab initio protein structure prediction. *Proteins: Struct. Funct. Genet.* **23**, 431–445.

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its

use in protein side-chain positioning. *Nature,* **356**, 539–542.
Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
Dunbrack, R. L., Jr & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains. *Nature Struct. Biol.* **1**, 334–340.
Durig, J. R. & Compton, D. A. C. (1979). Analysis of torsional spectra of molecules with two internal C3v rotors. 12. Low frequency vibrational spectra, methyl torsional potential function, and internal rotation of n-butane. *J. Phys. Chem.* **83**, 265–268.
Eisenmenger, F., Argos, P. & Abagyan, R. (1993). A method to configure protein side-chains from the main-chain trace in homology modeling. *J. Mol. Biol.* **231**, 849–860.
Faber, H. R. & Matthews, B. W. (1990). A mutated T4 lysozyme displays five different crystal conformations. *Nature,* **348**, 263–266.
Gelin, B. R. & Karplus, M. (1979). Side-chain torsional potentials: Effect of dipeptide, protein, and solvent environment. *Biochemistry.* **18**, 1256–1268.
Glantz, S. A. (1992). *Primer of Biostatistics,* McGraw-Hill, New York.
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.
Hwang, J. K. & Liao, W. F. (1995). Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.* **8**, 363–370.
James, M. N. G. & Sielecki, A. R. (1983). Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.* **163**, 299–361.
Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125**, 357–386.
Kishan, K. V., Zeelen, J. P., Noble, M. E., Borchert, T. V. & Wierenga, R. K. (1994). Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms. *Protein Sci.* **3**, 779–787.
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.
Kono, H. & Doi, J. (1994). Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins: Struct. Funct. Genet.* **19**, 244–255.
Kossiakoff, A. A., Randal, M., Guenot, J. & Eigenbrot, C. (1992). Variability of conformations at crystal contacts in BPTI represent true low-energy structures. *Proteins: Struct. Funct. Genet.* **14**, 65–74.
Lasters, I. & Desmet, J. (1993). The fuzzy-end elimination theorem: Correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.* **6**, 717–722.
Laughton, C. A. (1994). Prediction of protein side-chain conformations from local three-dimensional homology relationships. *J. Mol. Biol.* **235**, 1088–1097.
Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373–388.
Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.
McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295–310.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* **12**, 345–364.
Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham III, T. E., Ferguson, D. M., Seibel, G. L., Singh, U. C., Weiner, P. K. & Kollman, P. A. (1995). AMBER 4.1. University of California, San Francisco.
Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.
Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.
Ponder, J. W. & 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.
Ring, C. S. & Cohen, F. E. (1993). Modeling protein structures: construction and their applications. *FASEB J.* **7**, 783–790.
Sasisekharan, V. & Ponnuswamy, P. K. (1970). Backbone and side-chain conformations of amino acids and amino acid residues in peptides. *Biopolymers,* **9**, 1249–1256.
Sasisekharan, V. & Ponnuswamy, P. K. (1971). Studies on the conformation of amino acids. X. Conformations of norvalyl, leucyl and aromatic side groups in a dipeptide unit. *Biopolymers,* **10**, 583–592.
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.
Schulz, G. E. & Schirmer, R. H. (1979). Principles of Protein Structure. In *Springer Advanced Texts in Chemistry* (Cantor, C. R., ed.), Springer-Verlag, New York.
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.
Summers, N. L. & Karplus, M. (1989). Construction of side-chains in homology modeling. Application to the C-terminal lobe of rhizopuspepsin. *J. Mol. Biol.* **210**, 785–811.
Summers, N. L., Carlson, W. D. & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.* **196**, 175–198.
Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987a). Knowledge based modeling of homologous proteins, Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.
Sutcliffe, M. J., Hayes, F. R. & Blundell, T. L. (1987b). Knowledge based modeling of homologous proteins, Part II: rules for the conformations of substituted side-chains. *Protein Eng.* **1**, 385–392.
Swanson, E. (1994). In *PSSHOW Users Guide*1.9 edit.
Tanimura, R., Kidera, A. & Nakamura, H. (1994). Determinants of protein side-chain packing. *Protein Sci.* **3**, 2358–2365.
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.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1993). A critical comparison of search algorithms applied to the optimization of protein side-chain conformations. *J. Comput. Chem,* **14**, 790–798.

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

Vasquez, M. (1996). Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* **6**, 217–221.

Wiberg, K. B. & Murcko, M. A. (1988). Rotational barriers. 2. Energies of alkane rotamers. An examination of gauche interactions. *J. Am. Chem. Soc.* **110**, 8029–8038.

Wilson, C., Gregoret, L. M. & Agard, D. A. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996–1006.

Zhang, G., Kazanietz, M. G., Blumberg, P. M. & Hurley, J. H. (1995). Crystal structure of the cys2 activator-binding domain of protein kinase C delta in complex with phorbol ester. *Cell,* **81**, 917–924.

Edited by B. Honig

(Received 11 October 1996; received in revised form 10 November 1996; accepted 24 January 1997)