Protein Engineering vol.11 no.11 pp.991–997, 1998

Determinants of side chain conformational preferences in protein structures

Ram Samudrala[1,2] **and John Moult**[1,3]

[1]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850,
[2]Molecular and Cell Biology Program, University of Maryland at College Park, College Park, MD 20742, USA

[3]To whom correspondence should be addressed

A discriminatory function based on a statistical analysis of atomic contacts in protein structures is used for selecting side chain rotamers given a peptide main chain. The function allows us to rank different possible side chain conformations on the basis of contacts between side chain atoms and atoms in the environment. We compare the differences in constructing side chain conformations using contacts with only the local main chain, using the entire main chain, and by building pairs of side chains simultaneously with local main chain information. Using only the local main chain allows us to construct side chains with ~75% of the χ_1 angles within 30° of the experimental value, and an average side chain atom r.m.s.d. of 1.72 Å in a set of 10 proteins. The results of constructing side chains for the 10 proteins are compared with the results of other side chain building methods previously published. The comparison shows similar accuracies. An advantage of the present method is that it can be used to select a small number of likely side chain conformations for each residue, thus permitting limited combinatorial searches for building multiple protein side chains simultaneously.

Keywords: conditional probability/context-sensitivity/side chain rotamers

Introduction

Given a protein main chain conformation, constructing side chains by exploring all possible rotamer conformations simultaneously is a computationally intractable problem. Several approaches have been developed to reduce the number of possibilities. These include conformational searching using Monte Carlo and simulated annealing methods (Lee and Subbiah, 1991; Holm and Sander, 1992), using main chain dependent rotamer libraries to construct side chains (Dunbrack and Karplus, 1993), mean-field approaches (Koehl and Delarue, 1994), and matching local main chain coordinates to a database of side chain/main chain combinations (Kabsch *et al*., 1990; Wendolski and Salemme, 1992; Laughton, 1994).

The need to build side chains on a fixed main chain often arises in comparative modeling of protein structure, where a partial initial main chain conformation for the structure to be modeled (the target) is obtained from copying the main chain coordinates of parts of a related experimentally determined structure (the parent) (Greer, 1990; Mosimann *et al*., 1995). A comparison of the target and parent sequences is used to determine equivalent residue positions for which the main chain in the parent can be copied over to the main chain of the target. High homology main chain regions are not identical in conformation to the target structure, but can be quite similar (Mosimann *et al*., 1995; Martin *et al*., 1997).

We introduce a method that will reduce the number of conformational choices for a side chain based on a given environment, such as the local main chain. We use a conditional probability based discriminatory function (Samudrala and Moult, 1998a) to estimate the likelihood of a side chain conformation being correct. These probabilities are used to rank the different side chain conformations sampled using a discrete rotamer library. We perform an analysis of the accuracy of side chain construction considering only the local main chain (up to nine residues), using the entire main chain of the protein, and building side chains in a pairwise manner. We investigate the change in accuracy as the environment used for the construction of side chains is made more approximate. We evaluate the effect of the rotamer library approximation, and compare our results to other side chain building methods. We illustrate how side chain construction using only the local main chain can be combined with other search techniques to explore the conformational space of multiple protein side chains in the context of comparative modeling.

Methods

Description of discriminatory functions

Our objective here is to evaluate the correctness of a given side chain conformation in different environments. To do this, we use an all-atom distance dependent conditional probability-based discriminatory function to calculate the probability of observing a correct structure or substructure given the distances between pairs of atoms. A full description can be found in Samudrala and Moult (1998a). Briefly, the required probabilities are compiled by counting frequencies between pairs of atom types in a database of 265 experimental protein structures. All non-hydrogen atoms are considered, and the description of the atoms is residue specific, i.e. the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. We divide the distances observed into 1.0 Å bins ranging from 3.0 to 20.0 Å. Contacts between atom types in the 0.0–3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins. For observations of distances between pairs of atoms between the atoms of a side chain and the main chain of that residue, a separate table of frequencies is compiled using 18 1.0 Å bins ranging from 0.0 to 18.0 Å.

We compile tables of scores s proportional to the negative log conditional probability that we are observing a native conformation given an interatomic distance d for all possible pairs of the 167 atom types, a and b, for the 18 distance ranges, $P(C|d_{ab})$:

$$s(d_{ab}) = -\ln \frac{P(d_{ab}|C)}{P(d_{ab})} \alpha - \ln P(C|d_{ab}) \qquad (1)$$

where $P(d_{ab}|C)$ is the probability of observing a distance d

between atom types *a* and *b* in a correct structure, and $P(d_{ab})$ is the probability of observing such a distance in any structure, correct or incorrect. The required ratios $P(d_{ab}|C)/P(d_{ab})$ are compiled for all combinations of the 167 atom types for the 18 distance bins as follows:

$$\frac{P(d_{ab}|C)}{P(d_{ab})} = \frac{N(d_{ab})/\Sigma_d N(d_{ab})}{\Sigma_{ab} N(d_{ab})/\Sigma_d \Sigma_{ab} N(d_{ab})} \tag{2}$$

where $N(d_{ab})$ is the number of observations of atom types *a* and *b* in a particular distance bin *d*, $\Sigma_d N(d_{ab})$ is the number of *a*–*b* contacts observed for all distance bins, $\Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types *a* and *b* in a particular distance bin *d*, and $\Sigma_d \Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types *a* and *b* summed over all the distance bins *d*. No intra-residue distances are included in the summation.

The tables of scores are compiled from a set of non-homologous (less than 30% sequence identity between any proteins in the set) high-resolution (less than 3.0 Å) X-ray structures (Orengo,C., Michie,A., Jones,S., Swindells,M., Jones,D. and Thorton,J. (1993) <http://www.biochem.ucl.ac.uk/bsm/cath/>).

Given a set of *n* inter-residue distances between atoms *i* in a side chain and atoms *j* in the environment, and *m* intra-residue distances between atoms *k* in a side chain and atoms *l* in that residue's main chain, we calculate the score *S*(side chain) proportional to the negative log conditional probability that side chain conformation is native-like, using the expression:

$$S(\text{side chain}) = \sum_{ij}^{n} s(d_{ab}^{ij}) + \sum_{kl}^{m} s(d_{ab}^{kl}) \tag{3}$$

Description of rotamer library

Table I describes the main chain independent rotamer library used to sample the side chain conformations. For each rotamer, up to three χ angle values are defined. The library values were chosen by observing the preferences of side chains to be in discrete rotamer value bins in a database of protein structures. Other similar libraries have been developed (Ponder and Richards, 1987).

Table I. Main chain conformation independent rotamer library used to sample side chain conformations

Residue	Rotamer	Angle 1 (°)	Angle 2 (°)	Angle 3 (°)
C	χ_1	60.0	180.0	300.0
D	χ_1	60.0	180.0	300.0
D	χ_2	0.0	90.0	
E	χ_1	60.0	180.0	300.0
E	χ_2	60.0	180.0	300.0
E	χ_3	0.0	90.0	
F	χ_1	60.0	180.0	300.0
F	χ_2	0.0	90.0	
H	χ_1	60.0	180.0	300.0
H	χ_2	60.0	180.0	300.0
I	χ_1	60.0	180.0	300.0
I	χ_2	60.0	180.0	300.0
K	χ_1	60.0	180.0	300.0
K	χ_2	60.0	180.0	300.0
K	χ_3	60.0	180.0	300.0
K	χ_4	60.0	180.0	300.0
L	χ_1	60.0	180.0	300.0
L	χ_2	60.0	180.0	300.0
M	χ_1	60.0	180.0	300.0
M	χ_2	60.0	180.0	300.0
M	χ_3	60.0	180.0	300.0
N	χ_1	60.0	180.0	300.0
N	χ_2	60.0	180.0	300.0
Q	χ_1	60.0	180.0	300.0
Q	χ_2	60.0	180.0	300.0
Q	χ_3	60.0	180.0	300.0
R	χ_1	60.0	180.0	300.0
R	χ_2	60.0	180.0	300.0
R	χ_3	60.0	180.0	300.0
R	χ_4	60.0	180.0	300.0
S	χ_1	60.0	180.0	300.0
T	χ_1	60.0	180.0	300.0
V	χ_1	60.0	180.0	300.0
W	χ_1	60.0	180.0	300.0
W	χ_2	0.0	9.0	27.0
Y	χ_1	60.0	180.0	300.0
Y	χ_2	0.0	90.0	

Selection of the protein structures for testing side chain building

To build a test set of proteins we first obtained a list of 487 proteins with amino acid sequences less than 25% identical to each other, using the PDB SELECT tool (Hobohm *et al.*, 1992). From this set, all structures determined using NMR methods, all structures determined using X-ray crystallography having a resolution greater than 1.50 Å or an R-factor greater than 0.20, and all structures that were used in the compilation of the conditional probabilities for the atom type preferences were eliminated. Table II gives the details of the fifteen structures that were selected.

Generation of side chain conformations using only main chain information

All possible side chain conformations for each residue (excluding alanine, glycine and proline) were explored. The top scoring five conformations, based on the interactions between atoms in the side chain and the main chain, were selected. There are two sets of conformations: one based on interactions with only the local main chain (up to ± four residues, total of nine), and the other on interactions to the entire main chain.

Generation of side chain conformations in a pairwise manner

For each pair of residues with at least one interatomic contact within a distance of 6.0 Å, all combinations of side chain conformations are explored (excluding any pairs containing glycine, or proline residues). The pair of conformations with the best score, evaluated by summing the probabilities of the interactions between atoms of each of the two side chains with their respective local main chains (up to ± four residues, total of nine), and the probabilities of interactions between the atoms of the two side chains, were recorded. For each residue, the top scoring conformation from all the pairs it participates in was selected.

Evaluation of side chain construction

Selected conformations of each residue were compared with the corresponding experimental conformations. All the χ angles for a given side chain must agree with the experimentally observed values (i.e. *all* the χ angles for a given side chain must be within ± 60° or ± 45° of the corresponding experimental values depending on the residue type) in order for a side chain conformation to be considered correct. We do not consider proline residues in the evaluation. The 60° cut-off allows us to distinguish between an effect of the limited

Table II. List of proteins used to test side chain construction

Protein PDB code	Number of residues	Resolution (Å)	R-factor	Name
1bab-B	146	1.50	0.16	hemoglobin (human)
1cbn	46	0.83	0.11	crambin
1ccr	111	1.50	0.19	cytochrome C
1cus	197	1.25	0.16	cutinase
1pmy	123	1.50	0.20	pseudoazurin (cupredoxin)
1ptx	64	1.30	0.15	scorpion toxin II
1wfb-A	37	1.50	0.18	antifreeze protein isoform Hplc6
1xnb	185	1.49	0.17	xylanase
1xso-A	150	1.49	0.10	Cu, Zn superoxide dismutase
2end	137	1.45	0.16	endonuclease V
2hbg	147	1.50	0.13	hemoglobin (bloodworm)
2ihl	129	1.40	0.17	lysozyme (Japanese quail)
3sdh-A	145	1.40	0.16	hemoglobin I
2sga	181	1.50	0.13	proteinase A
9rnt	104	1.50	0.14	ribonuclease T1

The proteins were selected based on high resolution (⩽1.5 Å) and uniqueness (less than 25% sequence identity between any pair) and are not used in the compilation of the discriminatory function.

Table III. List of proteins used to compare accuracy of side chain construction with that of other methods

Protein PDB code	Number of residues	Resolution (Å)	R-factor	Name
1crn	46	1.5	0.11	crambin
1ctf	68	1.7	0.17	L7/L12 ribosomal protein
1lzl	130	1.5	0.18	lysozyme (human)
3apr	325	1.8	0.15	rhizopuspesin
2cro	65	2.4	0.20	λ cro repressor
3app	323	1.8	0.14	pencillopepsin
3tln	316	1.6	0.21	thermolysin
3fxn	138	1.9	0.21	flavodoxin
5pti	58	1.0	0.20	pancreatic trypsin inhibitor
7rsa	124	1.3	0.15	ribonuclease A

These proteins have been used previously to test side chain building in at least two of the methods described previously (Lee and Subbiah, 1991; Holm and Sander, 1992; Dunbrack and Karplus, 1993; Laughton, 1994).

rotamer library size versus a failure of the scoring function. We use the 30° cut-off as the strictest possible criterion for comparison with other methods.

Comparison to other methods

The accuracy of side chain conformations determined by local main chain contacts was compared with that obtained by other methods. Accuracy is assessed using all residues in a set of 10 structures that have been used by others to build side chains. We compare our method to those of Dunbrack and Karplus (1993), Holm and Sander (1992) Laughton (1994), and Lee and Subbiah (1991), by calculating the percentage of incorrect χ_1 angles, using a 30° cut-off (excluding alanine, glycine and proline residues), and the root mean squared deviation (r.m.s.d.) of the side chain atoms (including the C_β atom) between the built side chain conformation and the experimental conformations. These criteria were selected with the intent of being able to compare the approach described here with the largest number of other methods.

Details of the test set are given in Table III. In cases where different methods have used the same protein but with a different PDB structure [for example Lee and Subbiah (1991) have used 1rn3 instead of 7rsa for Ribonuclease A], we test our method using the PDB structure used by the method (other than our own) that gives the best results for that protein.

The discriminatory functions were recompiled removing all the proteins and homologs for which side chains were constructed.

The methods we choose are representative of the diverse set of methods available for side chain construction: Dunbrack and Karplus (1993) use a main chain dependent library of side chain rotamers to construct initial side chain conformations, and then use a minimization scheme to reorient side-chains that conflict with the main chain or other side chains.

Holm and Sander (1992) use a Monte Carlo algorithm together with the rotamer library of Tuffery *et al*. (1991) and simulated annealing with a simple potential energy function to optimize the packing of side chains on a given main chain.

Laughton compares the local environments of each side chain conformation to be built to a database of local environments for the same side chain type constructed from an analysis of protein structures. The database description consists of a list of C_α coordinates and residue type for each residue in the protein that has at least one atom within 4.0 Å of a side chain atom of the residue of interest. Side chain conformations that match the local environment criteria the best are input to a Monte Carlo procedure to give a final structure (Laughton, 1994).

Lee and Subbiah (1991) apply a simulated annealing algo-

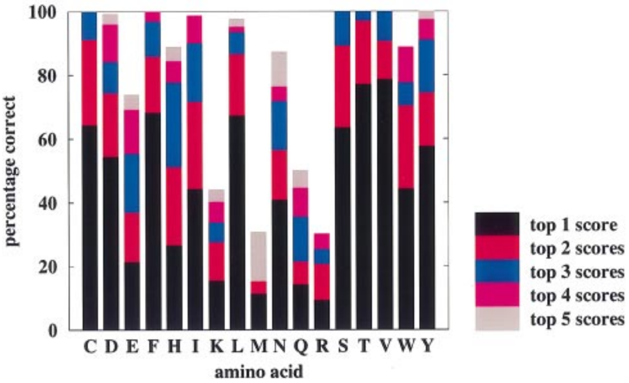

Fig. 1. Results of building side chain conformations for 17 amino acid types using the local main chain interactions only. The different coloured bars represent the percentage of side chains for which the correct conformation is in the top scoring one, two, three, four and five conformations. A conformation is considered correct if the rotamers for all the χ angles agree with those in the experimental structure (usually three rotamers per χ angle). As expected, the longer the side chain, the poorer the accuracy. Many side chains approach 100% accuracy when the five best scoring conformations are considered.

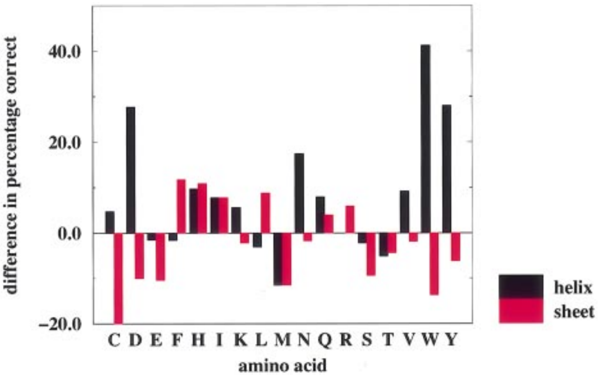

Fig. 2. Differences in accuracy of side chain construction for residues in α and β second structure relative to general accuracy. A positive bar indicates that a residue was built more accurately in the corresponding secondary structure than it was in general. Only the best scoring conformations for each side chain were used for this evaluation. Overall, the restrictions imposed by a helical main chain lead to improved accuracy.

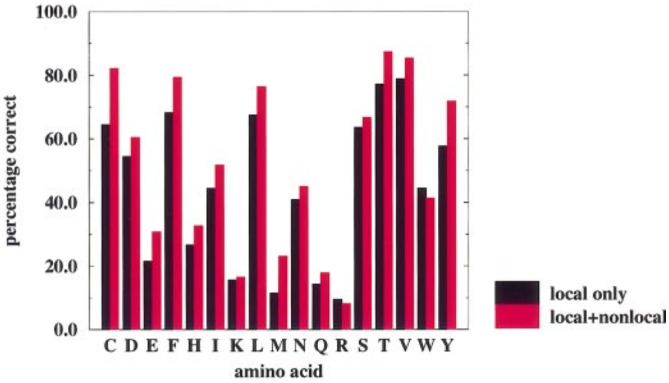

Fig. 3. Comparison of side chain construction using only local main chain interactions and that plus nonlocal interactions. Only the best scoring conformations are used for this evaluation. On average, accuracy improves by about 6% by adding the nonlocal information.

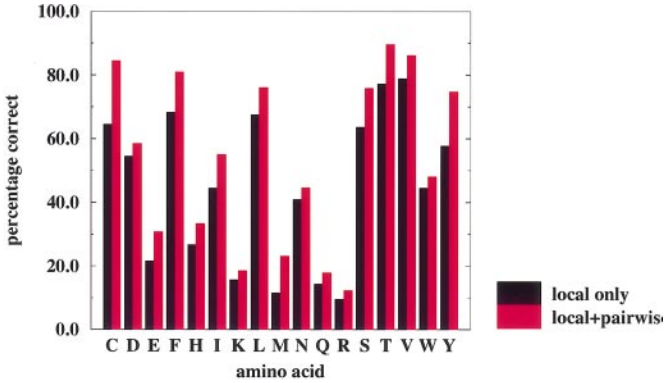

Fig. 4. Comparison of side chain construction using only local main chain interactions and that plus pairwise interactions. The side chain conformations with the best score using only the local main chain (± four residues, total of nine), and the side chain conformations in a pair of interacting side chains with the best score evaluated using both the local main chain and pairwise contacts for each residue, are used. The comparison is made for the 15 proteins in the test set. On average, accuracy is improved by about 10% by adding the pairwise information, with the largest gains in the most accurate cases.

rithm to the optimization of side chain packing interactions using a simple van der Waals potential function.

Effect of rotamer library approximation

Since we sample only up to three angles per rotamer (Table I), it is possible that the accuracy of our results using the percentage incorrect measure with a 30° cut-off or the side chain atom r.m.s.d. is limited by the non-discrete χ values in experimental structures. To investigate this effect, we find the closest rotamer library value for each side chain angle in the 10 experimental structures (Table III) and generate conformations using these values. We use these modified experimental structures to evaluate the limit of the accuracy of our rotamer library approximation.

Results

Accuracy of side chain construction for different residue types using only local main chain interactions

Figure 1 shows the percentage of side chains for which the correct conformation is in the set of the top scoring one, two, three, four or five conformations, using only local main chain information for the 17 different amino acids. The average overall accuracy is 51.9, 67.8, 78.5, 83.3 and 85.5% for each of these categories. Figure 2 shows the difference in the percentage accuracies using only the local main chain in cases where the residue adopts a α-helix or β-sheet secondary structure as classified by the program DSSP (Kabsch and Sander, 1983) and percentage accuracies regardless of secondary structure. The data are for all the residues in the 15 structures in Table II.

The average accuracy over all residue types for the single best scoring conformation is 52.6% for α-helix, 42.2% for β-sheet and 42.0% for residues not in α-helix or β-sheet. The average accuracy independent of secondary structure type is 44.8%. From Figures 1 and 2, it is evident that certain residues are more easily built for particular secondary structures. For example, phenylalanine in an α-helix is constructed accurately 66.6% of the time, whereas in a β-sheet the accuracy is 80.0%. Conversely, valine in an α-helix is constructed with 88.0% accuracy, whereas in a β-sheet, the accuracy is 76.9%. Some of the more dramatic differences include tryptophan (85.7% in α-helix, 30.7% in β-sheet), and aspartic acid (82.1% in α-helix, 44.4% in β-sheet). The side chains that are most difficult to build are the ones with the most χ angles and

Table IV. Comparison of side chain construction accuracy using the method of this work with four other previously published methods

Name of protein	PDB code	$\chi_1 > 30°$ (%)	Side chain atom r.m.s.d. (Å)	Dunbrack and Karplus (%)	Holm and Sander (%/Å)	Laughton (Å)	Lee and Subbiah (Å)
Crambin	1crn	13	1.40	8	–	1.43	1.65
L7/L12 ribosomal protein	1ctf	28	1.69	–	19/1.7	1.59	1.86
Lysozyme	1lzl	24	1.97	23	12/1.6	2.22	1.62
λ cro repressor	2cro	34	2.29	–	43/2.3	–	2.39
Pencillopepsin	3app	19	1.20	–	19/1.4	1.22	–
Rhizopuspesin	3apr	15	1.44	18	16/1.4	–	–
Flavodoxin	3fxn	37	1.76	–	39/1.9	1.96	1.90
Thermolysin	3tln	23	1.62	26	23/1.7	1.72	–
Trypsin inhibitor	5pti	21	1.73	15	22/1.9	2.61	1.49
Ribonuclease A	7rsa	33	2.02	21	21/1.8	2.02	1.86

The percentage error in χ_1 angles (using a 30° cut-off) excluding proline residues, and the side chain atom r.m.s.d. (including the C_β atom) are given. For the method of Dunbrack and Karplus, we list the percentage error in the χ_1 as given in (Dunbrack and Karplus, 1993), which includes prolines and uses a 40° cut-off; for Holm and Sander, we list the percentage error in the χ_1 angles (which includes prolines and uses a 30° cut-off) and the side chain atom r.m.s.d. as given in (Holm and Sander, 992); for Laughton and Lee and Subbiah, the side chain atom r.m.s.d., as listed in (Laughton, 1994) and (Lee and Subbiah, 1991) respectively, is given. All the methods produce similar results.

therefore the most degrees of freedom, such as glutamic acid, lysine, methionine, glutamine and arginine.

Accuracy of side chain construction including interactions with the entire main chain

The average overall accuracy of side chains constructed considering interactions with the entire main chain for the fifteen proteins in the test set is 57.8%, when the best scoring side chain is selected, an improvement of 5.9% compared with including only local interactions (Figure 3).

Accuracy of side chain construction using local and pairwise information

Figure 4 shows the results of adding residue pairwise interactions to local main chain interactions. There is an average improvement of about 10% in the accuracy of the side chain construction. This is a somewhat larger gain than adding interactions with the entire main chain for construction of side chains.

The residues built with a percentage accuracy of more than 80% with local and pairwise information are generally the ones with highest percentage accuracy when pairs of side chain conformations are evaluated simultaneously: phenylalanine–threonine, valine–threonine and cysteine–threonine have the largest pairwise percentage accuracies (of 75.0, 78.1 and 90.0%) among all pairs of residue types. Use of pairwise information produces the greatest improvement for the single residue accuracy of these four residues, compared with using only the local main chain information.

Comparison to other methods

Table IV shows the results of using our method on the set of 10 proteins for which side chains have also been constructed by other methods. The measures used for comparison are the percentage error in the χ_1 angles (i.e. the percentage of built conformations where the deviation in the χ_1 angle is greater than 30°) and the side chain atom r.m.s.d. (including the C_β atom) for all the residues in the protein (excluding alanine, glycine and proline residues).

Overall, with our method, five of the proteins have the lowest, or one of the lowest, percent error in the χ_1 angles, and six of the proteins have the lowest side chain atom r.m.s.d. Different methods have used slightly different criteria for calculating the percentage error in the χ_1 angle and the side chain r.m.s.d. We compare the performance of our method to each of those methods, taking into account the individual criteria used.

Dunbrack and Karplus (1993) used a cut-off of 40° for measuring the error in the χ_1 torsion angles and include proline residues in their calculation of the percentage of χ_1 angles correctly constructed. Taking the larger cut-off and including proline residues, the results for lysozyme and pancreatic trypsin inhibitor using our method are identical to theirs. For two of the proteins (rizopuspepsin and thermolysin), the percentage error is lower, and in two cases (ribonuclease A and crambin), the percentage error is higher.

Holm and Sander (1992) use a cut-off of 30° for the χ_1 torsion angle and include proline residues, although this does not change the relative performance of the two methods. In four cases the percentage error in the χ_1 angles is lower with our method, in two cases the percentage error is the same, and in three cases, it is worse. Holm and Sander (1992) also include the side chain atom r.m.s.d. (including the C_β atom) to a single digit precision. In four cases, our method produces lower side chain atom r.m.s.d. In three cases, the r.m.s.d. are about the same, and in the remaining two cases, the r.m.s.d. are worse.

Comparing our results to the side chain atom r.m.s.d. provided by Laughton (1994) for the eight structures whose side chain conformations are constructed, in six cases the r.m.s.d. are better, in one case the chain atom r.m.s.d. are identical, and we have a higher r.m.s.d. in only one case.

Lee and Subbiah (1991) produce higher r.m.s.d. than the method described here for four out of seven structures for which the side chain atom r.m.s.d. can be compared.

Effect of rotamer library approximation

Table V shows the effect of using the approximate rotamer library to sample side chain conformations. For each structure, side chain conformations with rotamer library values that are the closest to the experimental rotamer values are generated and compared with the experimental structure. The average percentage of side chains with χ_1 errors greater than 30° is 5.4%, and the average side chain r.m.s.d. error is 0.92 Å. The largest percentage error for χ_1 angles (of 12.1%) and the largest side chain atom r.m.s.d. (1.17 Å) is observed for 3fxn. 1crn is the structure with the lowest percentage error (0%) and the lowest side chain r.m.s.d. (0.70 Å). These values

Table V. Effect of using a discrete rotamer library approximation to sample side chain conformations

PDB code	Number of χ_1 rotamers	$\chi_1 > 30°$ (%)	Side chain r.m.s.d.
1crn	32	0.0	0.70
1ctf	46	6.5	0.88
1lzl	103	6.8	1.03
2cro	52	5.7	1.06
3app	247	3.2	0.88
3apr	243	3.2	0.86
3fxn	115	12.1	1.17
3tln	244	9.0	0.99
5pti	42	4.7	0.87
7rsa	105	2.8	0.83

The number of χ_1 angles considered, the percentage incorrect χ_1 angles (using a 30° cut-off) and the side chain atom r.m.s.d. (including the C_β atom) is given. These values represent the maximum accuracy a method can achieve using our rotamer library.

represent the maximum accuracy the methods described here can achieve.

Effect of experimental uncertainty

The percentage accuracies may be influenced by interatomic contacts between neighboring molecules in the crystals and local disorder. We rebuilt side chains using the local main chain for all the 15 proteins in the test set, excluding side chains having one or more atoms with a temperature factor greater than 30.0 Å^2. We performed a similar separate test excluding side chains having one or more atoms involved in an interatomic crystallographic contact of less than 4.0 Å to a neighboring molecule. Excluding side chains by each of these filters does not significantly change the results (less than 3% average percentage accuracy improvement for the fifteen proteins in both tests).

Discussion

Local main chain is most influential in determining side chain conformation

We find that the local main chain information alone is the most important factor for selecting the correct side chain conformation. Using this information, an average percentage accuracy of ~75% can be achieved in χ_1 angles (Table IV) and 52% for the entire side chain being correct (Figure 1), when the top scoring conformation is considered. More significantly, when the five best scoring conformations are considered, the correct complete side chain conformation is selected 82% of the time on average (Figure 1).

Other factors improve accuracy further: using the entire experimental structure main chain as the environment to determine side chain conformations increases accuracy by about 6% (Figure 3), and including the effect of the single most influential side chain with the local main chain information improves the accuracy by about 10% (Figure 4).

Since non-local main chain makes few contacts with a side chain, it is not surprising the effect is weak. However, the limited improvement from including the pairwise side chain interactions is not so expected.

Different chain building methods have similar accuracy

It is not straightforward to compare different methods because they have different goals and use different criteria for accuracy. Since there is often insufficient detail provided, we have tried to make our criteria as rigorous as possible and handle exceptions on a case-by-case basis (see the Results section). The method described here, using only the local main chain, compares favorably to the other methods (Lee and Subbiah, 1991; Holm and Sander, 1992; Dunbrack and Karplus, 1993; Laughton, 1994) published in the literature. All the four methods chosen for comparison in turn compare their methods to other methods and produce similar or slightly better results.

The methods used are very different yet produce similar levels of accuracy. The similarity in results using different methods, some of which are highly computer intensive (Lee and Subbiah, 1991), and some that require only a few seconds for a protein of any size (Dunbrack and Karplus, 1993), suggests that it is not too difficult to reproduce the correct side chain conformations with the experimental main chain to an average percentage accuracy of ~75% in the χ_1 angles (see Table IV). It appears that the large fraction of χ_1 angles in proteins are robustly determined by the environment and it does not matter much which method is used to determine them. However, accuracy for full side chains is much lower, and getting above this level of accuracy is more difficult and all the methods are equally ineffective.

Effect of rotamer library approximation

We find that limiting the number of rotamers per χ angle to three does not drastically affect the maximum possible accuracy of the method. Table V shows the limit of what the most accurate chain construction method can achieve given the rotamer library (Table I) for the set of 10 proteins, with an average percentage error of 5.4% for χ_1 angles (using a 30° cut-off).

Effect of secondary structure and residue type on side chain construction

The accuracy of side chain building generally depends on the secondary structure adopted by the local main chain (McGregor *et al.*, 1987; Dunbrack and Karplus, 1994). In our case, the average percentage accuracy for individual amino acids based on secondary structure type is 52.6 and 42.2% for α-helix and β-sheet secondary structures.

The accuracy of side chain building also depends on the residue type (Dunbrack and Karplus, 1993, 1994). It is perhaps relatively easy to select the right rotamer conformation in the case of a side chain with a single χ_1 angle with one degree of freedom (such as valine) with three possible values, compared with a side chain with four degrees of freedom with a total of $3^4 = 81$ possible values (such as lysine): random selection will yield percentage accuracies of 33.3 and 1.2% respectively for the two side chain types. However, even when comparing residues with similar degrees of freedom, ignoring secondary structure of the residue, there are differences in the percentage accuracy (Figure 1): isoleucine has a percentage accuracy of 44.1%, whereas leucine has an accuracy of 67.2%. Serine has an accuracy of 63.6% whereas threonine and valine have accuracies of 77.2 and 78.8% respectively.

Considering the effect of the combination of residue type and secondary structure of the main chain on side chain building also leads to interesting observations (Figures 1 and 2: isoleucine has an identical percentage accuracy of 52.5% in both α-helices and β-sheets, whereas leucine has a percentage accuracy of 63.9 and 75.8% in α-helices and β-sheets respectively. Threonine has a similar percentage accuracy in both α-helices and β-sheets (72.0 and 72.8%), but serine has an accuracy of 61.3% in α-helices and an accuracy of 54.2% in β-sheets.

Some of the observations are consistent with our understanding of the geometry of side chains and the geometry of secondary structure main chain (Dunbrack and Karplus, 1994). The side chains for most amino acid types are built more accurately in α-helices than in β-sheets. Exceptions are histidine, isoleucine, threonine and arginine, where the percentage accuracies are similar, and leucine and phenylalanine, where the accuracy is better in sheet than in helix. Presumably, the generally higher accuracy in α-helices is because the main chain conformation in helix regions reduces the number of degrees of freedom a residue side chain conformation can explore (Creamer and Rose, 1992, 1994).

Building side chains in a realistic modeling situation

Side chain building methods have generally been evaluated by re-building side chain conformations on an experimental structure main chain. However, it is very likely that in approximate environments, the side construction methods tested in idealized environments will not perform as well. For example, in a comparative modeling scenario, the main chain is approximate (~1.0 Å r.m.s.d.) and sometimes incorrect (>3.0 Å r.m.s.d.) even when there is a high (>50%) degree of local sequence identity between pairs of homologous structures (Samudrala *et al*., 1995; Samudrala and Moult, 1997). Chung and Subbiah have shown that as the main chain r.m.s.d. between homologous proteins rises to above 2.0 Å the average percentage accuracy for χ_1 angles goes down from 85 to 25% (using a 40° cut-off) for buried residues (Chung and Subbiah, 1995).

The increase in error in side chain construction with increasingly approximate main chain is because main chain and side chain conformations are intimately interconnected (Samudrala *et al*., 1995). A proper treatment of the problem of interconnectedness in protein structures would require the variation of the conformation of the side chains and the main chains simultaneously (Samudrala and Moult, 1997 and 1998b).

Using the set most likely side chain conformations to build full structures

Although the accuracy of the best scoring full side chain conformations is low, considering the top five best scoring ones produces a correct conformation more than 80% of the time. Further, the time taken to calculate the top five scoring conformations is only a few seconds even for large proteins. We have utilized this approach in a graph-theoretic clique finding method for selecting the best combination of side chain conformations, considering up to six conformations per residue, and including a set of 15–30 residues, as well as some main chain information. Details of this method are given in Samudrala and Moult (1998b) and its use in *bona fide* prediction is described in Samudrala and Moult (1997).

Application of this side chain construction approach using other discriminatory functions

To our knowledge, this is the first time a knowledge-based discriminatory function is used to select the most plausible side chain conformations using only the local main chain information. This approach is not limited to the function we use in this paper and is a means of testing the ability of other published scoring functions (Wallqvist *et al*., 1995; DeBolt and Skolnick, 1996; Subramaniam *et al*., 1996; Melo and Feytmans, 1997; Zhang *et al*., 1997) to build side chain conformations.

Availability of the software

The software to construct side chains using the approaches described in this paper is available via <http://www.ram.org/computing/ramp/ramp.html>.

Acknowledgements

Thanks to Maureen Perotto for assistance in preparing the plots. This work was supported in part by a Life Technologies Fellowship to Ram Samudrala and by NIH grant GM41034 to John Moult.

References

Chung,S. and Subbiah,S. (1995) *Protein Sci*., **4**, 2300–2309.

Creamer,T. and Rose,G. (1992) *Proc. Natl Acad. Sci. USA*, **13**, 5937–5941.

Creamer,T. and Rose,G. (1994) *Proteins Struct. Funct. Genet*., **19**, 85–97.

DeBolt,S. and Skolnick,J. (1996) *Protein Engng*, **9**, 637–655.

Dunbrack,R. and Karplus,M. (1993) *J. Mol. Biol*., **230**, 543–574.

Dunbrack,R. and Karplus,M. (1994) *Nature Struct. Biol*., **1**, 334–340.

Greer,J. (1990) *Proteins Struct. Funct. Genet*., **7**, 317–334.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci*., **1**, 409–417.

Holm,L. and Sander,C. (1992) *Proteins Struct. Funct. Genet*., **14**, 213–223.

Kabsch,W., Mannherz,H., Suck,D., Pai,E. and Holmes,K. (190) *Nature*, **347**, 37–44.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Koehl,P. and Delarue,M. (1994) *J. Mol. Biol*., **239**, 249–275.

Laughton,C. (1994) *J. Mol. Biol*., **235**, 1088–1097.

Lee,C. and Subbiah,S. (1991) *J. Mol. Biol*., **217**, 373–388.

Martin,A., MacArthur,M. and Thornton,J. (1997) *Proteins Struct. Funct. Genet*., in press.

McGregor,M., Islam,S. and Sternberg,M. (1987) *J. Mol. Biol*., **198**, 295–310.

Melo,F. and Feytmans,E. (1997) *J. Mol. Biol*., **267**, 207–222.

Mosimann,S., Meleshko,R. and James,M. (1995) *Proteins Struct. Funct. Genet*., **23**, 301–317.

Ponder,J. and Richards,F. (1987) *J. Mol. Biol*., **193**, 775–791.

Samudrala,R. and Moult,J. (1998a) *J. Mol. Biol*., **275**, 895–916.

Samudrala,R. and Moult,J. (1998b) *J. Mol. Biol*., **279**, 287–302.

Samudrala,R. and Moult,J. (1997) *Proteins Struct. Funct. Genet*., **29S**, 43–49.

Samudrala,R., Pedersen,J., Zhou,H., Luo,R., Fidelis,K. and Moult,J. (1995) *Proteins Struct. Funct. Genet*., **23**, 327–336.

Subramaniam,S., Tcheng,D.K. and Fenton,J. (1996) A knowledge-based method for protein structure refinement and prediction. In States,D., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology*, pp. 218–229.

Tuffery,P., Etchebest,C., Hazout,S. and Lavery,R. (1991) *J. Mol. Biol*., **217**, 373–388.

Wallqvist,A., Jernigan,R. and Covell,D. (1995) *Protein Sci*., **4**, 1881–1903.

Wendolski,J. and Salemme,F. (1992) *J. Mol. Graph*., **10**, 124–127.

Zhang,C., Vasmatzis,G., Cornette,J. and DeLisi,C. (1997) *J. Mol. Biol*., **267**, 707–726.

Received February 13, 1998; revised May 2, 1998; accepted May 6, 1998