J. Mol. Biol. (1993) 229, 996–1006

Modeling Side-chain Conformation for Homologous Proteins Using an Energy-based Rotamer Search

Charles Wilson[1]†, Lydia M. Gregoret[2]‡ and David A. Agard[1]

[1]*Howard Hughes Medical Institute*
Graduate Group in Biophysics and
Department of Biochemistry and Biophysics and
[2]*Department of Pharmaceutical Chemistry*
University of California at San Francisco
San Francisco, CA 94143-0448, U.S.A.

(Received 26 April 1991; accepted 8 October 1992)

We have developed a computational method for accurately predicting the conformation of side-chain atoms when building a protein structure from a known homologous structure. A library of rotamers is used to model the side-chains, allowing an average of five to six different conformations per residue. Local sites of adjacent side-chains are defined throughout the protein, and all combinations of side-chain rotamers are evaluated within each site using a molecular mechanics force field enhanced by the inclusion of a solvation term. At each site, the lowest energy combination of side-chains is identified and added onto the fixed protein backbone. A series of test cases using the refined X-ray structure of α-lytic protease has shown that: (1) the force field can correctly predict up to 90% of side-chain rotamers; (2) the assumption of side-chain rotamer geometry is usually a very good approximation; and (3) the complete combinatorial conformation search is able to overcome local minima and identify the lowest energy rotamer set for the protein in the absence of a starting bias to the correct structure. Tests with several pairs of homologous proteins have shown that the algorithm is quite successful at predicting side-chain conformation even when the protein backbone used to generate side-chain positions deviates from the correct conformation. The root-mean-square (r.m.s.) deviation of predicted side-chain atoms rises from 1·31 Å (average r.m.s.d. 0·73 Å) in a test case with the correct backbone to only 2·68 Å (1·95 Å average r.m.s.d.) in a test case with $< 35\%$ homology. The high accuracy of this method suggests that it may be a useful automated tool for modeling protein structure.

Keywords: homology modeling; side-chain rotamers; force fields; combinatorial search

1. Introduction

Several different approaches have been developed to predict the structure of a protein using its amino acid sequence. The most successful class of prediction techniques uses the known structure of an homologous protein as a starting point for modeling the unknown structure (Blundell *et al.*, 1987). As the number of protein crystal structures grows, it becomes increasingly likely that an unsolved protein will have some homology to a previously solved structure. In the future, therefore, it may be possible to predict the tertiary structure of many proteins using the database of solved structures. The development of an accurate, homology-based modeling algorithm is thus likely to have general use in solving the protein folding problem.

There are three major aspects to the problem of homology-based modeling: (1) constructing the sequence alignment between a pair of homologous proteins; (2) generating the backbone conformation of loops and other regions that are not conserved between the two proteins; and (3) predicting the conformation of side-chains, in both conserved and non-conserved regions of the structure. The current

† Present address: Department of Molecular Biology, Wellman 9, Massachusetts General Hospital, Boston, MA 02114, U.S.A.

‡ Present address: Department of Biology, MIT, Cambridge, MA 02139, U.S.A.

0022-2836/93/040996-11 $08.00/0

work deals only with the last part of this problem; modeling side-chain conformation.

There is no generally accepted method for predicting side-chain conformation, although several approaches have been developed. The methods may be divided into two broad categories: rule-based methods and conformation searching methods. Rule-based methods rely on the observation that topologically equivalent side-chains in homologous structures generally adopt the same torsion angles (Summers *et al.*, 1987; Chothia *et al.*, 1989; Summers & Karplus, 1989). However, there are known cases in which identical side-chains in homologous structures adopt different conformations (Summers *et al.*, 1987). Conformational searching approaches iteratively change the conformation of one side-chain at a time and score its conformation subject to a potential energy function (Snow & Amzel, 1986; Bruccoleri & Karplus, 1987; Novotny *et al.*, 1988). The method of Schiffer *et al.* (1990) has extended upon these methods by including a cycle of energy minimization prior to scoring. These algorithms are usually limited by the presence of multiple energy minima and so far solvation effects have not been taken into account. As a consequence, surface residues are generally modeled poorly. Furthermore, most conformational search methods developed to date are non-combinatorial in that they attempt to optimize only a single side-chain at a time. Recent work by Lee & Subbiah (1991) has shown that the core residue conformation can be more accurately predicted using a Monte Carlo procedure that simultaneously optimizes the complete set of side-chains for a protein.

Our approach is based on the observation by Ponder & Richards (1987) that side-chain conformations can be modeled using a relatively small library of idealized "rotamers". Instead of considering the full conformational space theoretically accessible to a side-chain, one of a small number of low energy conformations (typically 5 to 6 per residue type) can be used to describe the side-chain. In trying to model the conformation of side-chains, one need only specify which side-chain rotamer will be observed, rather than independently predict the co-ordinates of each atom. Because proteins are well packed, the conformation adopted by one side-chain can influence the conformations of neighboring residues. This suggests that a successful side-chain prediction algorithm must be multidimensional, simultaneously considering the combinatorial conformation space of several adjacent residues.

In a previous study, we have applied the rotamer representation of conformation space in combination with an approximate free energy force field to calculate the effects of mutagenesis on enzyme substrate specificity (Wilson *et al.*, 1991). Rotamers were used to model the three enzyme binding pocket residues and the P_1 side-chain of the substrate for mutants of the enzyme α-lytic protease. Surprisingly good agreement between calculated and experimental binding energies was obtained for a diverse body of enzyme kinetic data ($r > 0{\cdot}85$, average energy error = 0·7 kcal/mol for 42 mutant enzyme–substrate combinations; 1 cal = 4·184 J). Crystallographic studies of the mutant proteases complexed with boronic acid peptide inhibitors (transition state-like analogs for the substrates) showed that the lowest energy combination of binding pocket side-chain rotamers was generally the same as the set of rotamers observed in the X-ray structure. This observation suggested that the rotamer model could be used in a more general way to predict side-chain conformation given a fixed protein backbone.

Here we apply this computational method to the side-chain conformation prediction problem. While several other rotamer approaches have recently been published (Holm & Sander, 1991; Tuffery *et al.*, 1991; Desmet *et al.*, 1992), none has considered true homology modeling cases in which the structure of one protein is known and one is predicting that of a related protein (all papers have focussed on predicting side-chain conformation given the correct protein backbone co-ordinates). It thus remains unclear how well rotamers can be used to model side-chains when the protein backbone co-ordinates cannot be accurately known. To evaluate our method we deliberately choose test cases that mirror actual homology modeling situations: only the sequence of the "unknown" protein and the three-dimensional structure of an homologous protein are assumed to be known. Starting with the peptide backbone atom co-ordinates of the homologous structure (ignoring unconserved loops) and the correct sequence alignment, the appropriate side-chains corresponding to the sequence of the "unknown" protein are added. These are placed initially in arbitrary conformations. Sites are defined iteratively throughout the model and an attempt is made to optimize the local side-chain packing within each site subject to the energy function. Analysis of the structures before and after application of the above algorithm suggests that it is remarkably good at predicting side-chain conformation, especially for pairs of proteins with sequence identity.

2. Methods

(a) *Algorithm outline*

When modeling by homology, one is attempting to predict the structure of one protein, the unknown, using the experimentally determined structure of another homologous protein, the "template", as a starting point. Our algorithm, outlined in Fig. 1, produces a model for the unknown protein that consists of the backbone of the template protein with the appropriate unknown protein side-chain type added at each position. Co-ordinates for the unknown protein side-chain atoms are calculated outwards from the template protein backbone C^{α} atoms, assuming ideal bond lengths and bond angles. Side-chains are modeled as rotamers, and as such the dihedral angles between side-chain atoms (provided for by the Ponder–Richards rotamer library) correspond to low-energy staggered conformations observed in the database of protein crystal structures. Before generating a starting

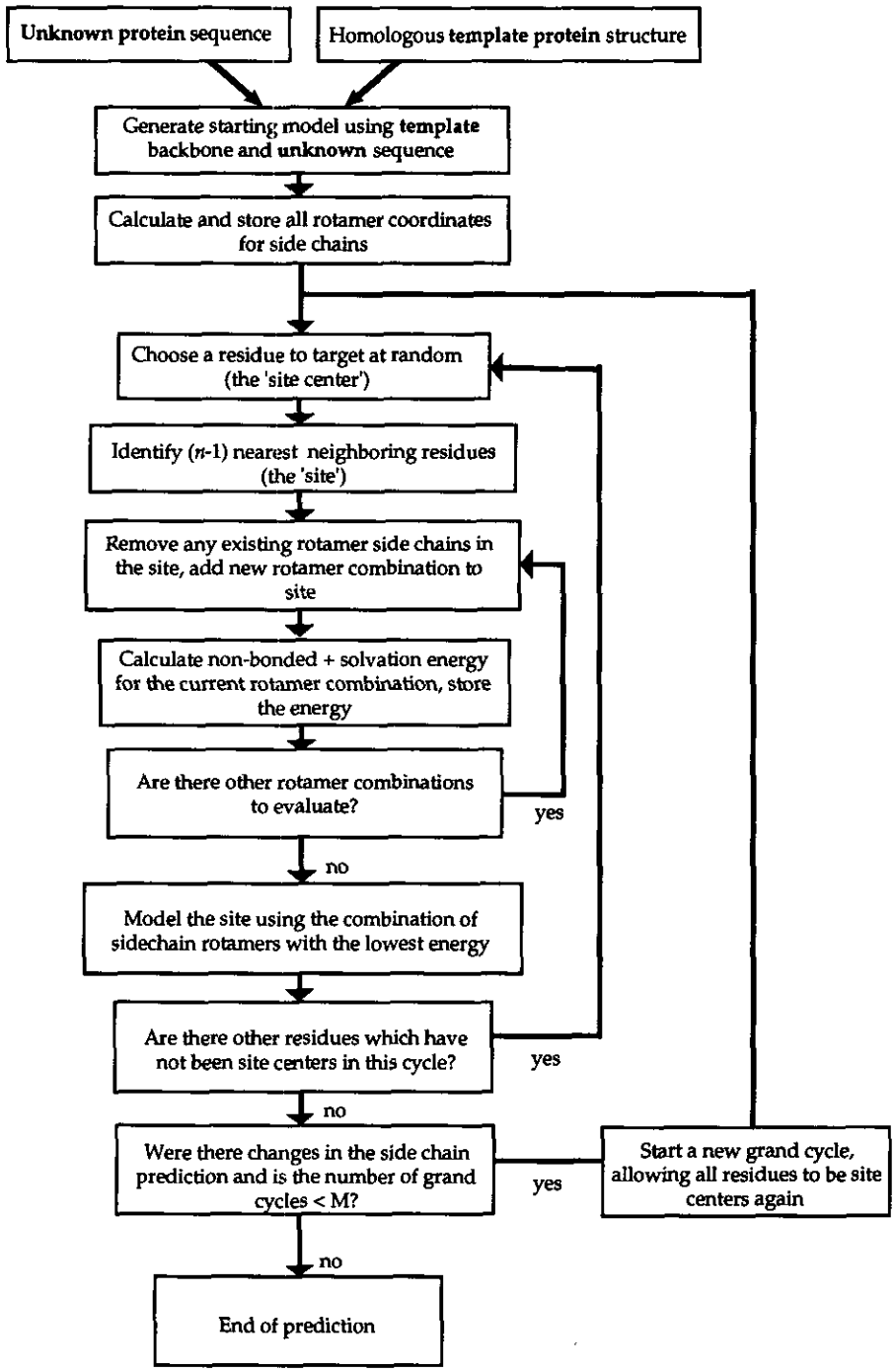

Figure 1. Algorithm for predicting side-chain conformation. Each step is described in detail in the text.

model, the co-ordinates for all rotamers that do not make significant bad contacts with the main-chain atoms are calculated and stored for use in the combinatorial searching. After generating the rotamers, the following procedure is applied (see Fig. 1). (1) A site center (one of the N amino acid residues in the model) is chosen at random. (2) A packing unit of n residues is identified and contains the site center side-chain and the $(n-1)$ other residues whose side-chain centroids are closest to that of the site center. The existing side-chains for these residues are deleted. (3) For the n residues in the site, all possible combinations of rotamers are tested, and for each rotamer combination, an approximate free energy is calculated. (4) After testing all combinations in the site, the set of side-chain rotamers that has the lowest calculated free energy is added to the model. (5) Steps (1) to (4) are repeated N times using a different central amino acid (randomly chosen from the remaining residues) until all the residues in the model have been used as site centers (adjacent sites may overlap and thus a given amino acid side-chain may be deleted and added back several times within a single cycle of refinement). (6) Steps (1) to (5) are repeated M times (using a different random order of site centers) or until the predicted side-chain conformations do not change from one cycle to the next. For the test cases reported here, each site included 5 residues ($n = 5$).

Repeated cycles did not converge for all residues because local sites overlap and in several cases, the rotamer chosen by the algorithm depended on which other residues were being considered simultaneously. For all of the test cases, therefore, the procedure was terminated after having cycled through the entire protein 3 times ($M = 3$). Increasing the size of the sites from 5 to 7 residues improved the results marginally but dramatically increased the computation time.

Details on the force field used to calculate the energy of each conformation are described in a previous paper (Wilson *et al.*, 1991). Since rotamers are used for the side-chains, the internal geometry of the atoms is idealized and thus the energetic costs of altering bond lengths, bond angles and dihedral angles can be neglected. The force field consequently has only two terms: (1) non-bonded interactions between atom pairs, and (2) the change in solvation energy upon exposing atoms to solvent. In our previous calculations of the effects of mutagenesis on enzyme substrate specificity, both the non-bonded and solvation terms were required to accurately reproduce experimentally determined binding energies. The optimal weights for the 2 terms found for the α-lytic protease calculations have been used for these studies ($w_{\text{Non-Bonded}} = 0{\cdot}031$, $w_{\text{Solvation}} = 1{\cdot}98$). The non-bonded terms (including electrostatics, hydrogen-bonding and van der Waals' energies) are calculated using the AMBER molecular mechanics force field (with a distance-dependent dielectric, $\varepsilon = r$, for the electrostatic term: Weiner *et al.*, 1984). Non-bonded interactions greater than 100 kcal/mol are truncated to this maximal value.

The effects of solvation are treated using a model similar to that of Eisenberg & McLachlan (1986). In their formalism, each atom is assigned an atomic solvation parameter (ASP†); multiplying an atoms' solvent-accessible surface area by its ASP directly gives the solvation energy for that atom. We have used the ASP approach but rather than calculating the precise solvent-accessible surface area for each atom (an extremely time-consuming computation), we use a grid method to estimate solvent accessibility. Each grid point on a 1·0 Å (1 Å = 0·1 nm) body-centered cubic lattice surrounding the macromolecule represents a pseudo-solvent molecule. For each conformation, a new list of allowed solvent positions is determined (solvent can be excluded from the grid by van der Waals' contacts with non-solvent atoms or by a lack of hydrogen-bonding partners). The number of grid points surrounding an atom occupied by solvent is proportional to the total atomic accessible surface area and thus to the solvation energy for the atom. There is no statistically significant difference between the accuracy of the original Eisenberg–McLachlan method and our grid-based method, as judged by the ability to fit amino acid transfer free energies (Wilson *et al.*, 1991).

(b) *Algorithm test cases*

The algorithm relies upon several approximations that are likely to introduce errors into the side-chain conformation prediction. To best isolate possible sources of error and to estimate the effect of various assumptions on the prediction, a series of test cases were constructed that incorporated the following changes in the modeling procedure. (1) The conformation of the side-chain that is present in the crystal structure was added to the set of rotamers considered for each residue (substituting for the Ponder–Richards library rotamer having the lowest r.m.s. deviation from the crystal structure side-chain). (2) All symmetry related atoms within 10 Å of any protein atom and the 2 crystallographically determined sulfate ions were included in the energy calculations (since crystal contacts may determine the conformation of surface side-chains). (3) Instead of combinatorially searching all rotamers at sites throughout the protein, a single residue was searched at a time and the lowest-energy conformation was identified. The true side-chain conformation was then added back to the protein so as not to bias the other positions. At the completion of the calculation, the predicted rotamers were then built onto the structure at each position. The 1st test case incorporated all 3 modifications, while test case no. 2 incorporated modifications (1) and (3) and test case no. 3 incorporated only modification (3).

(c) *Homology modeling test cases*

To evaluate the performance of our side-chain modeling algorithm, we constructed several "homology-built" models using pairs of known protein structures. An effort was made to select pairs of structures with a wide range of sequence similarities. The sequence identities between the pairs we tested (listed in Table 1) varied from 30% to 100%. The models were made by substituting the amino acid side-chains of the unknown structure onto the backbone of the template structure at all unambiguously equivalent positions. To determine which positions were structurally equivalent, the crystal structures were first superimposed manually using computer graphics, then refined automatically by finding all pairs of residues within 2·5 Å of each other. This alignment was then corrected by hand: in some cases where a loop in one structure had a similar conformation but was displaced by more than 2·5 Å from the analogous loop in the other structure, residues were still assigned to be equivalent. Loops having very different conformations and insertions were omitted from the models.

(d) *Model evaluation*

To evaluate the resulting side-chain conformations, the backbones of the true and model structures were superimposed by a least-squares fit method. The r.m.s. deviations of the side-chains and backbones of every residue in the model and true structures were then computed. We have tabulated both a side-chain-average r.m.s. deviation (averaging the r.m.s. deviation calculated for each residue), and an overall r.m.s. deviation (summed over all side-chain atoms). The average r.m.s. deviation (used by Novotny *et al.*, 1988) is generally smaller than the overall r.m.s. deviation since side-chains with fewer atoms tend to be more accurately modeled and their contribution is more highly weighted in the average deviation than in the overall deviation.

We also generated a best possible structure using the template structure backbone and the rotamer library. At each position, the rotamer having the lowest r.m.s. deviation from the side-chain in the true structure was selected. This "lowest co-ordinate error" structure is the best result we could hope to obtain with the algorithm since we allow only idealized rotamers for the side-chains. To score our model, we determined the fraction of the side-chains that had been assigned to the same rotamer as that in the lowest error structure.

† Abbreviations used: ASP, atomic solution parameter; r.m.s., root-mean-square; ALP, α-lytic protease; SGB, protease B from *S. griseus*.

Table 1

Test cases for side-chain conformation optimization

Model (template→unknown)	N_{unk}	N_{tmpl}	N_{mod}	Overall similarity (%)	Similarity of modeled regions (%)	Resol. unk (Å)	Resol. tmpl (Å)	Backbone r.m.s.d. (Å)
ALP→ALP	198	198	198	100·0	100·0	1·7	1·7	0·00
LZ1→LYZ	129	130	129	60·2	60·5	1·5	2·0	0·61
LBP→LIV	344	346	344	79·1	79·4	2·4	2·4	0·69
SGB→ALP	198	185	168	33·4	37·5	1·8	1·7	0·79
PTN→SGT	223	223	204	29·7	35·3	1·7	1·7	0·98

The "unknown" proteins were modeled using the "template" structures. N_{unk}, number of residues in the unknown structure. N_{tmpl}, number of residues in the template structure. N_{mod}, number of superimposable residues in the unknown and template structures that were modeled in the predicted structure. Overall similarity, percent sequence identity between the unknown and template structure determined using the sequence alignment method of Smith & Smith (1989). Similarity of modeled regions, percent sequence similarity for superimposable residues between unknown and template structures. Resol. unk, crystallographic resolution of the unknown structure. Resol. tmpl, crystallographic resolution of the template structure. Backbone r.m.s.d., root-mean-square deviation of backbone co-ordinates between the unknown and template structures for the residues modeled.

Structures used (Brookhaven PDB entry names in parentheses; Abola *et al.*, 1987): ALP, α-lytic protease (2ALP: Fujinaga *et al.*, 1985); SGB, protease B from *S. griseus* (3SGB: Read *et al.*, 1983); SGT, *S. griseus* trypsin (1SGT: Read & James, 1984); PTN, bovine trypsin (3PTN: Walter *et al.*, 1982); LYZ, hen egg-white lysozyme (6lyz: Diamond, 1974); LZ1, human lysozyme (1LZ1: Artymiuk & Blake, 1981); LIV, leucine/isoleucine/valine binding protein (2LIV: Sack *et al.*, 1989*a*); LBP, leucine binding protein (2LBP: Sack *et al.*, 1989*b*).

To determine whether our procedure actually improves the accuracy of the model, we also generated structures that had the most common rotamer assigned to each residue. The r.m.s. deviation of this unrefined first guess model from the true structure gives a baseline measure against which other models can be compared.

3. Results

(a) *Idealized tests with α-lytic protease*

Before applying the algorithm to homology model building, we first determined how accurately we could predict side-chain structure when the template protein backbone is completely correct (i.e. identical to the unknown protein backbone). Several tests (summarized in Table 2) were constructed to analyze different aspects of the prediction algorithm. In every case, we attempted to predict the side-chain structure for α-lytic protease, a 198 amino acid residue bacterial enzyme whose structure has been accurately determined at 1·7 Å resolution (Fujinaga *et al.*, 1985).

(i) *Test case no. 1 (force-field accuracy)*

In the first test case, the side-chains from the crystal structure of α-lytic protease were removed one at a time and then built back on to the protein backbone (details are provided in Methods). The only possible sources of error in this test case are the crystallographic co-ordinates for α-lytic protease and the force field used to estimate the free energy.

Of 142 residues with more than one rotamer (i.e. not glycine or alanine residues), the crystallographic side-chain rotamer was identified as the lowest energy rotamer in 126 cases (89% correct). The overall r.m.s. deviation between the lowest energy structure and the crystal structure was 0·59 Å (side-chain atoms only, see Table 2). While the majority of side-chains are modeled correctly, there are certainly a significant number of errors in this best case model.

Assuming that the incorrect predictions result from errors in the force field used to evaluate them, we hoped to better understand these problems by analyzing the characteristics of the poorly placed side-chains. Of the 16 residues that were not correctly predicted, 12 were exposed and four were buried. The bias towards exposed residues is not unexpected since there are strong packing constraints on buried residues, which do not exist at the surface. Surprisingly, all four incorrectly predicted buried residues were polar amino acids,

Table 2

Results of the α-lytic protease test cases

Test case name	Starting with crystal structure side-chains?	Include crystal contacts, counterions?	Library rotamer replaced by crystal structure side-chain?	Average r.m.s. deviation (Å)	Overall r.m.s. deviation (Å)	Fraction correct (correct/total)
Force-field accuracy	Yes	Yes	Yes	0·26±0·58	0·59	0·89(126/142)
Effect of crystal contacts	Yes	No	Yes	0·27±0·61	0·62	0·89(126/142)
Effect of rotamer approximation	Yes	No	No	0·68±0·85	1·21	0·82(116/142)
Standard algorithm	No	No	No	0·73±0·91	1·31	0·76(111/142)

Results for modeling the side-chains of α-lytic protease using the true backbone are shown. Average r.m.s. deviation, average root-mean-square deviation of non-alanine side-chains of the predicted structure from the true structure (unweighted by the number of atoms in each side-chain). Overall r.m.s. deviation, root-mean-square deviation of all side-chain atoms.

including three serine residues and one asparagine residue (Val40 was the only hydrophobic amino acid among the 16 incorrect residues). The incorrect residues include an unusually high number of serine (6) and asparagine (5) residues, perhaps indicating that uncharged hydrogen bonds may be treated improperly by the force field. The incorrectly predicted side-chains often form a set of hydrogen bonds that differ from those actually found in the crystal structure.

(ii) *Test case no. 2 (effect of crystal contacts)*

The conformation of many solvent-accessible residues may be determined in the crystal by contacts with symmetry related molecules. To test this possibility, we repeated the calculation but neglected to include symmetry related molecules and bound counter-ions. The effect appeared to be negligible as the overall r.m.s. deviation increased only slightly to 0·62 Å with no net change in the number of incorrectly modeled side-chains (Table 2). In all subsequent tests, symmetry related atoms and bound sulfate groups have been ignored.

(iii) *Test case no. 3 (effect of the rotamer approximation)*

In the first test case we replaced the rotamer in the library with the lowest r.m.s. deviation by the true side-chain at each site to make certain that errors in the prediction were not due to the assumption of idealized rotamer geometry. In the third test case, we used only the standard library rotamers to determine the effect of the rotamer approximation. Using the library rotamers, the overall r.m.s. deviation for side-chain atoms increased from 0·62 Å to 1·21 Å (116 of 142 side-chains (82%) were modeled correctly). The lowest error rotamer structure (calculated using the library rotamers) has an overall r.m.s. deviation of 0·71 Å; the predicted structure is thus only 0·5 Å worse than the best structure possible given the constraint of ideal rotamer geometry. Amino acid residues with long side-chains (lysine, arginine, glutamine) account for more than two-thirds of the residues that were initially predicted but are incorrectly placed after reverting to the standard rotamer library. This bias is not surprising since the rotamer approximation should be worst for those amino acids with many torsion angles.

(iv) *Test case no. 4 (standard algorithm)*

To test the ability of the algorithm to converge without the correct neighboring side-chains, the above test was repeated with a starting structure that was completely stripped of side-chains. Using the standard algorithm (with sites of 5 residues, cycling through the sequence 3 times, adding the lowest energy rotamer combination in each case), the number of correct side-chains plateaued at 111 residues (*versus* 116 residues in the previous test). This result indicates that the combinatorial conformation search converges well in the absence of a starting bias towards the correct structure.

The general conclusions of the α-lytic protease test cases (summarized in Table 2) are as follows. (1) The force field is able to correctly predict almost 90% of the observed side-chain conformations, with the incorrectly predicted side-chains lying largely at the surface and including a disproportionately high number of serine and asparagine residues. (2) Symmetry related atoms and bound counter-ions do not significantly affect the ability to predict the side-chain conformation observed in the crystal. (3) Using side-chain rotamers rather than the true side-chains prevents the correct prediction of approximately 7% of the residues. (4) By combinatorially searching local sites throughout the protein, it is possible to accurately predict most side-chains without a starting bias to the correct structure. With no errors in the protein backbone, but no starting information about side-chain conformations, the overall r.m.s. deviation for α-lytic protease side-chains was 1·31 Å. Similarly good results were obtained for a number of other proteins (Table 3), indicating the general utility of the algorithm.

Table 3

Predicting side-chain conformation using the correct peptide backbone

Protein	Lowest error rotamer structure r.m.s. deviation	Final overall r.m.s. deviation	Final buried residue r.m.s. deviation (% correct)	Final accessible residue r.m.s. deviation (% correct)
2ALP	0·71	1·31	0·93(88)	1·56(67)
1CRN	0·38	1·32	0·82(88)	1·41(66)
5PTI	0·67	1·67	1·76(69)	1·65(58)
1CTF	0·63	1·49	1·06(75)	1·60(34)

Side-chains were initially deleted from all proteins and then built on using the standard algorithm described in the text. Solvent accessibility was calculated using the method of Lee & Richards (1971) as implemented in the program ACCESS. Residues considered buried have less than 20% of their accessible surface areas exposed (relative to an extended tripeptide model). Labels correspond to the Protein Data Bank file containing the starting co-ordinates: 2ALP-α-lytic protease (Fujinaga *et al.*, 1985), 1CRN-crambin (Hendrickson & Teeter, 1981), 5PTI-bovine pancreatic trypsin inhibitor (Walter & Huber, 1983) and 1CTF-L7/L12 50 S ribosomal protein (Leijonmarck & Liljas, 1987).

Table 4
Homology modeling results

Structure	First guess average r.m.s.d.	First guess overall r.m.s.d.	Predicted average r.m.s.d.	Predicted overall r.m.s.d.	L.E. average r.m.s.d.	L.E. overall r.m.s.d.	Fraction correct buried	Fraction correct exposed	Fraction correct total
ALP→ALP	1·69±1·31	2·48	0·73±0·91	1·31	0·39±0·44	0·71	0·88 (58/66)	0·67 (52/78)	0·78
LZ1→LYZ	1·99±1·21	2·58	1·44±1·02	1·90	0·91±0·49	1·06	0·78 (32/41)	0·53 (34/64)	0·63
LBP→LIV	2·05±1·16	2·49	1·55±1·05	1·96	1·10±0·56	1·25	0·78 (96/123)	0·43 (61/143)	0·59
SGB→ALP	2·29±1·47	3·03	1·88±1·50	2·66	1·30±1·07	1·73	0·70 (37/53)	0·57 (39/68)	0·63
PTN→SGT	2·40±1·60	3·17	1·95±1·60	2·68	1·44±1·27	1·92	0·81 (54/67)	0·46 (38/83)	0·61

Results for homology modelling test cases. Abbreviations for the proteins are the same as in Table 1. L.E., lowest error structure.

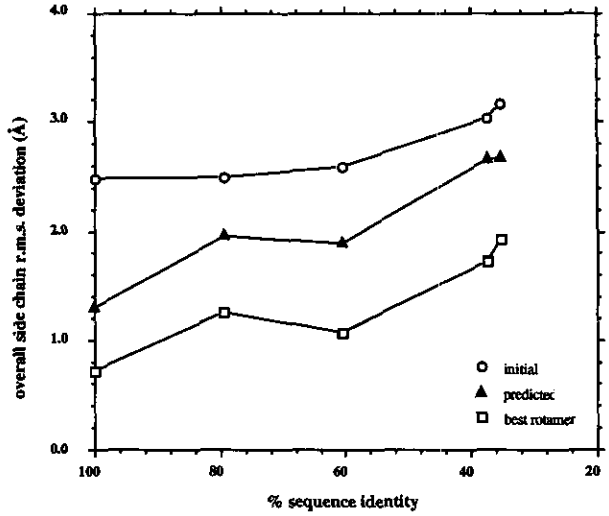

Figure 2. Accuracy of the side-chain prediction as a function of percent homology. The overall side-chain r.m.s. deviation is shown as a function of the percent homology between the unknown and predicted structures for the initial, predicted and best rotamer (lowest r.m.s. deviation) models.

(b) *Homology modeling*

With the accuracy of the force field and the rotamer assumption well tested, we have proceeded to use the algorithm to predict side-chain conformations for pairs of homologous proteins. These homology modeling tests differ from the α-lytic protease test cases in two respects: (1) the backbone used to calculate side-chain rotamers is not exactly correct (since the template structure backbone differs from the true structure backbone); and (2) the template peptide backbone is punctuated by gaps where non-homologous regions have not been modeled. Our results are summarized in Table 4. In every case, there is a significant improvement in the accuracy of the model following application of the algorithm. Not surprisingly, the ability to correctly predict side-chain conformation decreases as the deviation between the model backbone and true backbone increases. The improvement, as measured by r.m.s. deviation to the true structure or by the fraction of correctly predicted side-chains, drops approximately linearly with decreasing sequence identity (Figure 2).

As with the α-lytic protease test case, solvent-accessible residues are significantly harder to predict than buried residues. Figure 3 shows the predicted structure of hen egg-white lysozyme, with both hydrophobic core residues and some surface residues. While aromatic residues making up the core are all accurately positioned, exposed residues are often incorrect. The fraction of buried or solvent-accessible residues that are correctly placed for each test case are listed in Table 4. Increased errors at the surface are likely to be due to a number of phenomena including greater allowed conformational space (since there are fewer restricting

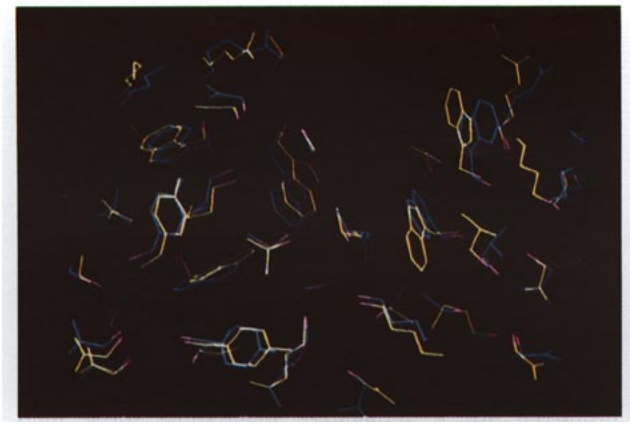

Figure 3. Comparison between the predicted and observed hen egg-white lysozyme structures. Side-chain and C^{α} atoms are shown for the predicted (blue) and true (yellow) structures (C^{α} atoms colored magenta). Residues in the hydrophobic core lie on the left-hand side while those on the right are generally somewhat solvent accessible.

adjacent residues), increased crystallographic errors in the surface residues and errors in the force field that effect electrostatic interactions more than van der Waals' interactions (given that hydrophilic residues predominate at the surface). Previous analysis of protein crystal structures has shown that surface residues have systematically higher temperature factors than buried residues (Alber *et al.*, 1987), indicating that their side-chain atoms are less well fixed in an energy minimum. This observation suggests that the energy differences between alternate conformations may be smaller at the surface and that slight errors in the force field should affect surface residues more than buried ones.

By comparing the predictions made for the ALP → ALP test and the SGB → ALP test, we can quantify the errors introduced by using the wrong backbone to predict side-chain conformations. Using the standard iterative procedure to place α-lytic protease side-chains on the stripped backbone of α-lytic protease, 78% of side-chains (111/142) are correctly predicted. This fraction drops to 63% (76/121) when the backbone of *Streptomyces griseus* protease B is used instead. The average r.m.s. deviation of side-chains also increases in going from the α-lytic protease backbone (0·73 Å) to the *S. griseus* protease B backbone (1·88 Å). This increase is higher than that observed for the backbone atoms (0·00 Å for ALP, 0·79 Å for SGB), suggesting that errors in the backbone positions adversely affect the choice of side-chain rotamer, beyond simply displacing the side-chain away from the correct position.

Figure 4 shows a representative case in which deviations in the backbone between a pair of homologous structures directly lead to an incorrect side-chain choice. The backbone atoms of the neighboring residues isoleucine 105 and tyrosine 237 in *S. griseus* protease B deviate by only approximately 0·4 Å relative to their equivalents in α-lytic protease (tryptophan 105 and tyrosine 238). These relatively

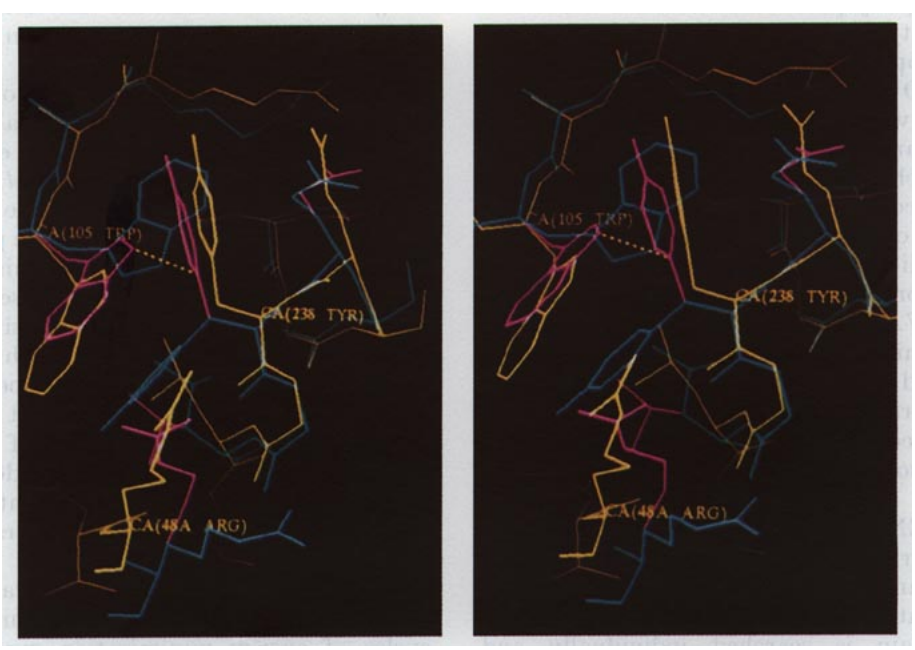

Figure 4. Errors in the *S. griseus* protease B→α-lytic protease prediction. The true (yellow) and predicted (blue) structures for α-lytic protease are shown (using *S. griseus* protease B as a backbone template). The best rotamers (those with the lowest r.m.s. deviation to the true structure) are shown in magenta. Several bad contacts between the best rotamers for Trp105 and Tyr238 (e.g. $N^{\varepsilon 1}$-105–CG-238 distance = 2·43 Å, dotted line) force an alternate set of rotamers to be chosen as the lowest energy conformation. The misplaced tyrosine 238 ring subsequently forces Arg48A to adopt an incorrect conformation. All 3 residues are correctly positioned when using the true α-lytic protease backbone to generate the side-chain rotamers (not shown).

small deviations alter the direction of the $C^{\alpha} \rightarrow C^{\beta}$ vectors sufficiently to cause a significant change in the positions of the calculated rotamers. The r.m.s. deviation of the lowest error rotamers at these positions from the true side-chains rises from 0·2 Å (using the α-lytic backbone) to 1·4 Å (using the *S. griseus* protease B backbone). More importantly, the SGB → ALP lowest error rotamer structure has several bad van der Waals' contacts between tryptophan 105 and tyrosine 238, causing this combination of rotamers to be ignored during the rotamer search. For instance, the separation between $N^{\varepsilon 1}$ of tryptophan 105 and C^{α} of tyrosine 238 drops from the close distance of 3·02 Å in the ALP → ALP lowest error rotamer structure to the bad contact distance of 2·43 Å in the SGB → ALP structure (Fig. 4). Whereas the lowest error rotamers for residues 105 and 238 are identified as the lowest energy combination in the ALP→ALP test, an alternate set of rotamers is chosen for the SGB→ALP case. The same is true of Arg48A, which lies adjacent to this pair of residues. While it is correctly placed for the ALP→ALP test case, the incorrectly positioned tyrosine 238 in the SGB→ALP test forces this arginine residue into an incorrect position.

4. Discussion

This work has shown that a combinatorial rotamer search directed by an approximate free energy calculation can be used to predict side-chain conformation in a homology modeling test. The fraction of properly placed side-chains is a function of the similarity between the pair of homologous structures, dropping from approximately 80% in the case of 100% identity, to approximately 60% for those tests with lower homology. By using rigid rotamers to coarsely sample conformation space and a grid approach to calculate solvent accessibility, the complete combinatorial search can be carried out extremely quickly. Starting with the backbone alone, the prediction of side-chain conformation for a 200 residue protein can be completed in less than five hours of VAX 8650 CPU time. Our algorithm provides a significant improvement, both in terms of accuracy and speed, over energy-based side-chain modeling algorithms that have been previously reported (Bruccoleri & Karplus, 1987; Schiffer *et al.*, 1990). The reasons for this improvement will now be considered.

The CONGEN program of Bruccoleri & Karplus (1987) uses a grid search over main-chain and side-chain torsion angles to model both loops and side-chain conformation. The conformation space of each added side-chain is searched individually and evaluated using the CHARMM force field. This molecular mechanics force field includes terms for covalently linked atom pairs (bond-stretching, bond angle bending, torsion angle rotation) and for non-bonded pairs (van der Waals' forces, electrostatics and hydrogen-bonding). After evaluating all staggered conformations, the lowest energy conformation is saved. This method can replace side-chains onto a structure with the correct backbone with an r.m.s. deviation of approximately 2·5 Å (averaged over side-chains, not including C^{β} atoms).

While conceptually similar to the approach we have described, there are several major differences between the two methods. In contrast to our program, CONGEN includes bonded-energies but ignores solvation effects in evaluating side-chains. Since our approach uses rotamers with idealized internal geometry, there is no need to consider the bonded terms. Work by Bruccoleri and others, however, has shown that the side-chain rotamers preferred by a molecular mechanics force field lacking solvation terms are biased towards those that have relatively unfavorable solvation energy (Novotny *et al.*, 1988; Schiffer *et al.*, 1990). By not taking into account solvation effects during the rotamer search, therefore, the CONGEN approach tends to incorrectly predict the conformation of polar surface side-chains (Novotny *et al.*, 1988).

A second difference between the two methods is in their approach to sampling conformation space. Whereas the CONGEN program can search an arbitrary number of rotamers for a single side-chain, our algorithm combinatorially tests a limited number of rotamers at each site for a cluster of adjacent residues. By simultaneously varying several side-chain conformations, energetic barriers to co-operative rearrangements can be surmounted. It is interesting to note that if our algorithm is applied using sites containing a single residue rather than five adjacent residues, an additional approximately 10% of the side-chains are not correctly predicted after the first cycle (data not shown).

Schiffer *et al.* (1990) describe a method for constructing side-chains that is closely related to the CONGEN approach. In this algorithm, staggered side-chain conformations are evaluated using the AMBER force field (Weiner *et al.*, 1984). A zone surrounding each targeted residue is subject to energy minimization to improve the packing around the altered side-chain. The final minimized energy for each side-chain orientation is used to determine which rotamer is adopted at each site. As with the CONGEN algorithm, this approach does not take into account solvation effects and does not combinatorially test adjacent side-chains. It does, however, have the significant advantage of allowing side-chains to deviate from their initial idealized rotamer geometry. In cases in which slight bad contacts exist between the library rotamers (e.g. Fig. 4), energy minimization should allow the contacting atoms to relax and thereby yield a more realistic energy estimate. Because several hundred thousand cycles of energy minimization must be done to complete a single cycle of side-chain optimization, this approach is extremely computer-intensive. It has currently been applied to only a subset of the residues in the one test case that has been reported (bovine → rat trypsin). As presently implemented, this method seems promising but it may require a

significant increase in computer-speed to be generally practicable.

Recent work by Lee & Subbiah (1991) has applied the technique of simulated annealing to the problem of predicting side-chain conformation. Their algorithm uses a force field containing only torsional and van der Waals' terms to evaluate any given combination of side-chains. Discrete steps of 10 degrees in the side-chain torsion angles are taken to sample different conformations. Rather than defining a limited site within the protein, the complete set of protein side-chains is simultaneously optimized, using an annealing Monte Carlo procedure to coarsely sample a wide section of conformation space. This procedure has been applied to a number of test cases in which the correct backbone is used but all side-chain atoms are initially deleted. While good results are obtained for buried residues (average r.m.s. deviation = 1·25 Å), errors in the surface residues are significantly higher, causing the average overall r.m.s. deviation of side-chain atoms to rise to 1·97 Å. Results obtained using our algorithm (Table 3) indicate that the use of a more complete force field (including electrostatic and solvation terms) significantly improves upon their predictions. Whereas buried side-chains are predicted with approximately equal accuracy by either the Lee & Subbiah algorithm or ours, our overall r.m.s. deviations are approximately 25% lower; indicating a significant improvement for surface residues. *A priori*, one might have expected the Lee & Subbiah algorithm to more accurately predict buried side-chain conformation since their method allows a much finer conformation search (36 conformers/torsion angle *versus* 5 to 6 rotamers/side-chain with our method). The fact that buried residues are predicted equally well by both methods indicates that the assumption of rotamer geometry for side-chains is usually a very good approximation.

We have demonstrated how small errors in the backbone co-ordinates can force the incorrect choice of side-chain rotamers. By restricting the conformation search using a rigid backbone, it is currently impossible to overcome these problems. It may be possible to significantly improve upon our results by carrying out constrained energy minimization or constrained molecular dynamics on the full structure between the cycles of side-chain optimization. If the side-chains are correctly modeled onto the template backbone, relaxation of the structure using energy minimization or molecular dynamics should allow backbone atoms to move towards their true co-ordinates, generating the "unknown" backbone. As the side-chain conformations gradually improve between cycles, the ability to minimize to the correct backbone should also increase. The potential benefits of this approach may be limited by the accuracy of the force field used to direct the energy minimization. In most simulations starting from a crystal structure, energy minimization causes r.m.s. shifts of approximately 0·5 to 1·0 Å. However, for side-chain modeling in cases of low homology this level of error is likely to be small enough to allow an improvement over the current fixed-backbone procedure. It may be possible to reduce the severity of errors in the force field by applying a constraining potential in either the minimization or dynamics calculations to minimize the deviations from the template backbone.

Funding for this research was provided by the Howard Hughes Medical Institute. C.W. was supported by a Fannie and John Hertz Foundation fellowship in applied physics. L.M.G. was supported by a predoctoral fellowship from the National Science Foundation. We thank Jay Ponder and Fred Richards for kindly supplying us with the program PROPAK.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Crystallographic Databases – Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Alber, T., Sun, D. P., Nye, J. A., Muchmore, D. C. & Matthews, B. W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, **26**, 3754–3758.

Artymiuk, P. J. & Blake, C. C. F. (1981). Refinement of human lysozyme at 1·5 angstroms resolution. Analysis of non-bonded and hydrogen-bond interactions. *J. Mol. Biol.* **152**, 737–762.

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 (London)*, **326**, 347–352.

Bruccoleri, R. E. & Karplus, M. (1987). Prediction of folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, **26**, 137–168.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D. & Tulip, W. R. (1989). Conformations of immunoglobulin hypervariable regions. *Nature (London)*, **342**, 877–883.

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London)*, **356**, 539–542.

Diamond, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* **82**, 371–391.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature (London)*, **319**, 199–203.

Fujinaga, M., Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1985). Refined structure of alpha-lytic protease at 1·7 angstroms resolution. Analysis of hydrogen bonding and solvent structure. *J. Mol. Biol.* **184**, 479–502.

Hendrickson, W. A. & Teeter, M. M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)*, **290**, 107–113.

Holm, L. & Sander C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218**, 183–194.

Lee, B. & Richards, F. M. (1971). The interpretation of

protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

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

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L7-L12 from *Escherichia coli* at 1·7 angstroms. *J. Mol. Biol.* **195**, 555–579.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Func. Genet.* **4**, 19–30.

Ponder, J. A. & Richards, F. M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequence for different structural classes. *J. Mol. Biol.* **193**, 775–791.

Read, R. J. & James, M. N. G. (1984). Critical comparison of comparative model building of *Streptomyces griseus* trypsin. *Biochemistry*, **23**, 6570–6575.

Read, R. J., Fujinaga, M., Sielecki, A. R. & James, M. N. G. (1983). Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1·8 angstroms resolution. *Biochemistry*, **22**, 4420–4433.

Sack, J. S., Saper, M. A. & Quiocho, F. A. (1989*a*). Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. *J. Mol. Biol.* **206**, 171–191.

Sack, J. S., Trakhanov, S. D., Tsigannik, I. H. & Quiocho, F. A. (1989*b*). Structure of the L-leucine-binding protein refined at 2·4 angstroms resolution and comparison with the leu/ile/val-binding protein. *J. Mol. Biol.* **206**, 193–207.

Schiffer, C. A., Caldwell, J. W., Kollman, P. & Stroud, R. M. (1990). Prediction of homologous protein structures based on conformational searches and energetics. *Proteins: Struct. Func. Genet.* **8**, 30–43.

Smith, R. F. & Smith, P. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *PNAS*, **87**, 118–122.

Snow, M. E. & Amzel, L. M. (1986). Calculating the three-dimensional changes in protein structure due to amino acid substitutions: The variable region of immunoglobulins. *Proteins: Struct. Func. Genet.* **1**, 267–279.

Summers, N. L. & Karplus, M. (1989). Construction of side-chains in homology modelling: 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.

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.

Walter, J. & Huber, R. (1983) Pancreatic trypsin inhibitor. A new crystal form and its analysis. *J. Mol. Biol.* **167**, 911–917.

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W. & Huber, R. (1982). On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography. *Acta Crystallogr. (sect. B)*, **38**, 1462–1472.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Amer. Chem. Soc.* **106**, 765–784.

Wilson, C., Mace, J. E. & Agard, D. A. (1991). A computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* **220**, 496–506.

Edited by B. Honig