J. Mol. Biol. (1991) **217**, 373–388

Prediction of Protein Side-chain Conformation by Packing Optimization

Christopher Lee† and S. Subbiah

Beckman Laboratories for Structural Biology
Department of Cell Biology
Stanford University Medical Center
Stanford, CA 94305, U.S.A.

(Received 22 June 1990; accepted 5 October 1990)

We have developed a rapid and completely automatic method for prediction of protein side-chain conformation, applying the simulated annealing algorithm to optimization of side-chain packing (van der Waals) interactions. The method directly attacks the combinatorial problem of simultaneously predicting many residues' conformation, solving in 8 to 12 hours problems for which the systematic search would require over 10^{300} central processing unit years. Over a test set of nine proteins ranging in size from 46 to 323 residues, the program's predictions for side-chain atoms had a root-mean-square (r.m.s.) deviation of 1·77 Å overall *versus* the native structures. More importantly, the predictions for core residues were especially accurate, with an r.m.s. value of 1·25 Å overall: 80 to 90% of the large hydrophobic side-chains dominating the internal core were correctly predicted, *versus* 30 to 40% for most current methods. The predictions' main errors were in surface residues poorly constrained by packing and small residues with greater steric freedom and hydrogen bonding interactions, which were not included in the program's potential function. van der Waals interactions appear to be the supreme determinant of the arrangement of side-chains in the core, enforcing a unique allowed packing that in every case so far examined matches the native structure.

1. Introduction

Protein folding naturally breaks down into two classes of degrees of freedom: the ϕ–ψ torsions, which determine the main-chain fold, and the χ torsions, which set the pattern of side-chain packing. These two sets of variables are closely coupled, because of the tremendous importance of side-chain packing for the stability of the overall fold. Many efforts at protein structure prediction have focused on prediction of side-chain conformation (given knowledge of the main-chain) as a less complex but nonetheless important subproblem of the protein folding problem (Richards, 1977; Blow, 1983; Janin *et al.*, 1978; Warme & Morgan, 1978*a*,*b*; Greer, 1981; James & Sielecki, 1983; Narayana & Argos, 1984; Lesk & Chothia, 1986; Ponder & Richards, 1987; Blundell *et al.*, 1987; Bruccoleri & Karplus, 1987; Summers *et al.*, 1987; Reid & Thornton, 1989; Summers & Karplus, 1989; Singh & Thornton, 1990). As many workers have pointed out, the ability to predict side-chain conformation accurately from the sequence and a model of the main-chain fold would be a useful tool for homology modeling (Delbaere, 1979; Blundell *et al.*, 1983, 1987; Chothia, 1984; Read *et al.*, 1984; Greer, 1985; McCormick *et al.*, 1985; Sibanda & Thornton, 1985; Chothia *et al.*, 1986; Cohen *et al.*, 1986; Chothia & Lesk, 1987; Cohen & Kuntz, 1987; McGregor *et al.*, 1987; Pearl & Taylor, 1987; Sutcliffe *et al.*, 1987*a*,*b*; Zvelebil *et al.*, 1987; Strynadka & James, 1988; Taylor, 1988; Reid & Thornton, 1989; Weber *et al.*, 1989*a*,*b*) and other applications such as the construction of protein models from X-ray diffraction data after the initial tracing of the main-chain (Jones & Thirup, 1986).

As with the larger problem of protein folding, the principal difficulty in making predictions of side-chain conformations is the enormous number of structural permutations possible for even small (5 to 10 residue) prediction problems. Considering side-chain torsions as rigid rotations divided into discrete 10° steps (so that each χ angle has $360°/10° = 36$ distinct possible states), a problem containing n torsions permutes to 36^n possible structures. For five residues with ten χ torsions this corresponds to $3{\cdot}7 \times 10^{15}$ conformations, or about 10^5 VAX 8650 CPU‡ years for a rapid energy calcu-

† Author to whom all correspondence should be addressed.

‡ Abbreviations used: CPU, central processing unit; r.m.s., root-mean-square.

0022–2836/91/020373–16 $03.00/0

lation algorithm performing the systematic search for a lowest energy conformation. On the face of it, simultaneous optimization of multiple side-chains seems an impossibly time-consuming operation (Reid & Thornton, 1989; Summers & Karplus, 1989).

Two different strategies have been used in attempts to get around this permutational obstacle. The first was to drastically reduce the number of conformations allowed for each residue type to only a few basic rotamers, as in the method described by Ponder & Richards (1987). In this formalism, each side-chain is only tested in three to seven fixed conformations, instead of being allowed to rotate freely; for a five residue zone each with five tested rotamers, this reduces the permutations to only 3125, allowing multiple sequences to be tested. This approach has been reported as a powerful method for enumerating the possible sequences that can be packed into a given region of a protein, and has indicated strict limitations imposed by packing on core sequences. Since this model still has an exponential dependence on the size of the problem (e.g. 100 residues with 5 tested rotamers each would require examining $7{\cdot}9 \times 10^{69}$ permutations) it is difficult to apply to large predictions.

An alternate approach has been simply to forgo the attempt to predict all side-chains simultaneously, opting instead to make predictions residue by residue, weighing the possible combinations intelligently, and gradually assigning side-chain conformations in order of most reliable to least reliable. This strategy has been examined in ground-breaking studies both using energy calculational methods (Gelin & Karplus, 1975, 1979; Bruccoleri & Karplus, 1987; Summers & Karplus, 1989), and knowledge-based human modelers (Reid & Thornton, 1989). Reid and Thornton, in particular, were able to predict side-chain conformations of flavodoxin with an overall r.m.s. of 2·41 Å (1 Å = 0·1 nm) starting from C^{α} co-ordinates alone, using computational methods to predict main-chain atoms, and manual examination and adjustment using computer graphics to predict the side-chains. The main strength of this approach, its ability to cut through the enormous permutations simply by ignoring them, unfortunately turns out to be its main weakness: it is not able to explore adequately the possible combinations of side-chain packings. In the protein core, conformations of neighboring side-chains are strongly coupled, such that an error in one residue often leads to errors in adjacent residues, propagating throughout the prediction (Reid & Thornton, 1989). In the flavodoxin prediction, initial misplacement of a core phenylalanine residue led to a variety of errors causing the internal core to be incorrectly packed. In such cases, it is the combination of side-chain conformations that must be considered, to determine the best conformation for each individual residue, rather than the other way around. Thus, the question of how to deal with the complexity of solving multiple residues' positions simultaneously appears to be an important unsolved problem, critical to accurate side-chain prediction.

We have sought to address this problem directly. Since coupling of side-chain positions is mediated primarily by the necessity of avoiding steric overlap, we decided to approach side-chain prediction as a problem of minimizing van der Waals packing interactions *via* simultaneous rigid rotations of all side-chains. We have thus sought to predict side-chain positions solely by finding ways that pack them together well. The main difficulty lies in finding a search strategy which can locate these good packings in a reasonably short amount of computing time, since the combinatorix of this problem is enormous even for relatively small prediction problems. As most observers have pointed out, the brute-force strategy of systematic search would require impossibly long computations (Reid & Thornton, 1989; Summers & Karplus, 1989). In contrast, the simulated annealing algorithm (Metropolis *et al.*, 1953) has been used with success for a variety of similar NP-complete optimization problems (Kirkpatrick *et al.*, 1983; Van Hemmen & Morgenstern, 1983; Brunger, 1988; Subbiah & Harrison, 1989), and allows approximate solution of such problems in a fraction of the computing time required for the systematic search. We have therefore applied simulated annealing to the side-chain packing problem, and have assessed its performance in predicting side-chain co-ordinates over a set of nine example proteins.

This strategy faces two distinct challenges. First, it must overcome the combinatorial complexity of this problem to find well-packed conformations. In densely packed protein cores, only a minute fraction of the total conformation space is free of unacceptable steric collisions (Richards, 1977; Ponder & Richards, 1987). Finding well-packed conformations is thus a significant computational challenge. Second, it is unclear that packing energy is a strong predictor of actual protein structure. Although each protein's actual structure is by necessity well-packed, there might be other good ways of packing the internal side-chains, some of which might be very different (Ponder & Richards, 1987; Lim & Sauer, 1989). If there were such a degeneracy of multiple good packings, packing alone would be a poor predictor of actual structure, and other interactions (such as electrostatics) would have to be brought into consideration. Indeed, even relatively sophisticated potential functions have been found to be poor predictors of actual structure (Novotny *et al.*, 1984). This issue is of interest because it touches on the basic question of the importance of packing constraints to determining the structure of proteins. For this reason, we have limited our prediction method to include *only* van der Waals interactions in its evaluation of side-chain conformations, so as to assess explicitly the degeneracy of good packings, and their usefulness as predictors of actual structure.

2. Method

The principal difficulty in the side-chain packing problem is the tremendous size of the solution space that must be searched. The key to overcoming this difficulty is the intuitively obvious fact that one need not visit every site in the solution space to locate the general regions of energy minima. If the energy function is sufficiently continuous and exhibits smooth trends throughout the solution space, minima can be located by traversing paths across the solution space, and using the observed trends to direct the search to just those regions where minima are likely to be found.

The simulated annealing algorithm exemplifies this approach, and has been used extensively in computer science to solve a wide variety of optimization problems (Metropolis *et al.*, 1953; Kirkpatrick *et al.*, 1983; Van Hemmen & Morgenstern, 1983; Brunger, 1988; Subbiah & Harrison, 1989; for a general description of the method, see Press *et al.*, 1986). It uses a random walk to traverse the solution space, with a bias towards minimal energy zones, controlled by the annealing temperature T. Specifically, at each stage of the random walk a small perturbation from the current position is randomly selected as a move, and the energy change (ΔE) associated with making this move calculated. If $\Delta E \leq 0$ then the move is accepted. If $\Delta E > 0$ then a random number is used to decide whether or not to accept the move, with an acceptance probability of:

$$p = e^{-\Delta E/T},$$

where T is the "annealing temperature".

In the annealing procedure this walk is started with T very large, so that $p \approx 1$ regardless of the value of ΔE; thus the walk is effectively unconstrained. Over the course of the walk, however, T is gradually reduced, causing the walk to become more and more strongly biased towards reducing the total energy. This algorithm corresponds to walking randomly over the energy landscape, gradually reducing the rate of escape from "valleys". This results in a walk continuously passing through segments of most of the major minima, gradually increasing the fraction of time spent in the deepest minima. Done slowly enough, this focusing procedure leads eventually to spending almost all the time walking around the region of the deepest minimum.

For this to work, the random walk must be able to traverse the solution space, i.e. travel from one extreme to the other along any dimension, rapidly and without impediment, so as to pass through all the major minima zones. Energetic barriers that block travel between minima prevent the walk from exploring the space sufficiently to locate the global minimum. The (12, 6)-potential ordinarily used for modelling van der Waals interactions has the form:

$$E = \varepsilon_0[(r_0/r)^{12} - 2(r_0/r)^6],$$

where ε_0 and r_0 are constant parameters describing, respectively, the depth of the energy well, and the equilibrium interatomic distance for the van der Waals interaction of a given pair of atoms (see Table 1 for values used), which becomes infinite as r tends to zero (for a general discussion of the Lennard–Jones potential, see Atkins, 1986). These infinite energy barriers would block simulated annealing from exploring the solution space, and have therefore been truncated to a maximum value of 7 kcal/mol (1 cal = 4·184 J) for each pairwise interaction (similar to the "soft atoms" described by Levitt, 1983*b*).

Table 1

Energetic parameters used for side-chain predictions

$Atom_1$	$Atom_2$	r_0(Å)	ε_0(kcal/mol)
C	C	4·315	0·0738
C	O	3·916	0·1168
C	N	4·058	0·1746
C	S	4·315	0·0738
O	O	3·553	0·1848
O	N	3·683	0·2763
O	S	3·916	0·1168
N	N	3·817	0·4132
N	S	4·058	0·1746
S	S	4·315	0·0738

This same problem, not traversing the solution space sufficiently to locate the global minimum, can also be caused by lowering the annealing temperature T too rapidly. If the temperature is reduced too quickly, preventing the walk from crossing the energy barriers between minima, the walk will most likely be trapped in a local minimum. In practice, T is lowered in discrete steps, so that both the number of random walk moves per T-step and the amount of temperature decrease per T-step determine the effective rate of cooling. The number of moves per step has been chosen to guarantee extensive and repeated traversal of the solution space during each T-step. Since the r.m.s. deviation of a random walk of n unitary steps is $\sqrt{n}$, the number of steps needed to generate an r.m.s. deviation d from the starting position is just d^2 (Van Kampen, 1981). Using 10 deg. increments and a 50% probability of moving a given torsion on each step, the number of steps necessary to generate on average a 180° deviation is $(180/10)^2/(0{\cdot}50) = 648$. We have found that 10,000 move-steps per T-step works well for this problem, giving around 15 traversals of the solution space per T-step. Following previous reports (Kirkpatrick *et al.*, 1983; Subbiah & Harrison, 1989), we decrease the annealing temperature by 2% each T-step.

Application of this method to the side-chain packing problem has been relatively straightforward. In each step all side-chain torsion angles are randomly moved −10°, 0° or +10° as rigid rotations, with all bond lengths and angles held to equilibrium values used previously in molecular dynamics simulations (given in Table 3 of Levitt, 1983*a*). Hydrogen atoms have not been included in the molecular representation; instead we use the slightly augmented van der Waals radii appropriate for the "united atom" representation commonly employed in molecular dynamics, such as the program ENCAD of M. Levitt (personal communication) (see Table 1). We have made no attempts to find parameters that optimize prediction accuracy, nor have we made any modifications of the parameters from their standard values.

To make simulated annealing function well for the side-chain packing problem, we have added several features that enhance its ability to seek minima and escape energetic "traps". To improve the program's ability to locate conformations in which all side-chains are well-packed, we developed an algorithm to identify individual residues that were well-packed. After every move, each residue's total van der Waals interaction with its surroundings was calculated. If this residue energy was less than a threshold good packing value, the residue was placed in refinement mode; for the next 1000 steps, its move probability was reduced 2-fold. In this way residues that are found to be

well-packed are gently constrained while the program seeks good conformations for those that are not. This method helps the walk approach the global minimum much faster. For prediction problems larger than about 5 residues, the simulated annealing walk typically became trapped in local minima at temperatures well above those needed to seek the global minimum. To help the program escape such traps, we altered the move algorithm to slowly increase its move size (from 10° to 20° to 30° etc.) whenever it became trapped (i.e. when the program failed to find a single acceptable move in over 100 consecutive steps), gradually enlarging the range of its allowed moves until it was able to find an acceptable move. The move size was then reset to normal.

To speed energy calculations during the annealing run, a variety of precalculations were performed. First, simple torsional potentials:

$$E_{\text{torsion}} = A \cos(3\chi),$$

$A = 1{\cdot}5$ kcal/mol, only for χ_1 and χ_2, and van der Waals interactions between side-chain atoms and main-chain atoms, which remain immobile throughout the walk, were precalculated and stored. To calculate interactions between side-chain atoms more efficiently during annealing, condensed lists of all pairwise interactions approaching close enough for significant van der Waals interaction were prepared beforehand. To assess the closest approach distance for each pair of side-chain atoms, every pair of side-chain atoms was "aimed" at one another by twisting the appropriate torsions to minimize the distance between them. If this approach distance was less than 5 Å, the pair was added to the list. During annealing, the total list of these possible pairwise interactions was scanned periodically, and only pairs that were less than 5 Å apart at that step were calculated during the 10 subsequent steps, thus further reducing the number of interactions that had to be calculated to just those making a significant contribution. Since the r.m.s. deviation for 10 steps of this random walk is $\sqrt{10} \times 0{\cdot}50 \times 10° \approx 20°$, the updating is repeated frequently enough to preserve the accuracy of the calculated energy.

To start the annealing, all side-chain torsions are set to random angles selected from a uniform probability distribution 0 to 360°, and an initial temperature chosen to ensure free movement over most energy barriers. Following previous reports (Kirkpatrick *et al.*, 1983), we select our starting temperature so as to give a move accept/reject ratio of 1/1, that is, the temperature at which 50% of the proposed moves are accepted (referred to hereafter as T_{50}). At this temperature, the average energy and standard deviation of the energy are similar to that of the entire solution space, and the walk frequently attains energy levels signifying energy maxima. To find this T_{50} value the program initially adjusts its temperature rapidly until the accepts fraction falls close to 50%; then annealing is begun. As the annealing temperature is gradually reduced and the walk locates minima, the lowest energy structures found are noted and stored. When the accepts fraction falls below 20%, the walk is terminated.

A typical annealing run generates about 2000 low energy (more than 3σ below the mean) conformations. To synthesize a single predicted structure from this aggregate set, a weighted average energy is calculated for each residue in each of its possible conformations, and its conformation with the lowest energy average over the set is taken as the predicted structure for the residue. The weighted average energy for a given conformation (specified by χ_1 and χ_2) is calculated according to:

$$E_{\text{av}}(\chi_1, \chi_2) = \sum E_i(\chi_1, \chi_2)\, e^{-E_i(\chi_1, \chi_2)/E_{\text{wt}}} / \sum e^{-E_i(\chi_1, \chi_2)/E_{\text{wt}}},$$

where E_{wt} is the weighting energy determining how selective the average should be. This is simply a Boltzman average, treating the set of reported structures as a canonical ensemble. For side-chains with 3 or more torsions, the best position for the χ_1 and χ_2 torsions is determined *via* the weighted average; once all residues have been solved out to χ_2 in this way, the remaining χ_3 and χ_4 torsions are solved simply by spinning them independently through their full range to find the lowest energy conformation for each residue.

(a) *Selection of residues for prediction*

Although the method was originally designed to predict side-chain conformations in local zones of 5 to 15 residues within a protein, we have found it useful for simultaneous prediction of whole proteins. For the results described in this paper, 9 proteins (crambin, 1crn (Hendrickson & Teeter, 1981); bovine pancreatic trypsin inhibitor, 5pti (Wlodawer *et al.*, 1984); the C-terminal domain of the ribosomal protein L7/L12, 1ctf (Leijonmarck & Liljas, 1987); human lysozyme, 1lz1 (Artymiuk & Blake, 1981); ribonuclease A, 1rn3 (Borkatoti *et al.*, 1982); 434cro, 2cro (Mondragon *et al.*, 1989); flavodoxin, 4fxn (Smith *et al.*, 1977); thermolysin, 3tln (Holmes & Matthews, 1982); and penicillopepsin, 2app (James & Sielecki, 1983)) from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) were chosen for their high-resolution, accurate structures, and for comparison with other predictive methods that have been applied on these proteins. The 7 smaller structures fell well within program memory limitations for simultaneous prediction of all side-chains possessing free χ torsions. In each of these cases the native co-ordinates for main-chain and C^β atoms, and all proline atoms, were used as the basis for the predictions, which were each generated in a single annealing run per protein. In order to keep our definition of side-chain r.m.s. deviation consistent with previous reports, we have included C^β atoms in the calculation of side-chain r.m.s. values; however, it should be noted that these atoms were drawn directly from the native co-ordinate set. In our experience, this approach produces side-chain r.m.s. deviations only very slightly lower than those obtained using C^β positions predicted from native main-chain co-ordinates. For side-chains with a rotational symmetry axis (Phe, Tyr, Asp, Glu), r.m.s. and torsion angle comparisons against the X-ray structure took symmetry into account, by explicitly applying the symmetry operator (in this case a 2-fold rotation axis) to each such residue and selecting the conformation with the lower error.

For each of the 2 larger proteins (app, tln), only the 100 or so most buried residues were predicted, due to memory limitations that prevented running the whole proteins. In both of these cases, only main-chain, C^β, and proline atoms were fed to the prediction; to avoid prejudicing the predictions, the co-ordinates for the remaining unpredicted surface side-chains were wholly deleted prior to annealing. This was done out of concern that the presence of the surface side-chains locked into their correct conformations would strongly bias the predictions towards the native structure. No attempt was made to compensate for their loss, and the empty space left inside the protein by their disappearance does not appear to have weakened the prediction. To select the buried residues for prediction, a simple algorithm which calculates the fraction of accessible surface area for each residue (similar to the method of

Connolly, 1985) was employed with a simple cutoff value that included about one-third of the residues in each protein. For purposes of comparison, residues in the other proteins of the set have been defined as "core" in Tables 2 and 3 if their accessible surface area fractions were below this same threshold value.

(b) *Computing time*

The prediction procedure is broken into 3 separate stages: setup, annealing and averaging. The setup program performs a variety of precalculations, generating parameter files needed to drive the annealing program. For prediction of all residues in flavodoxin (138 residues), setup took 30 min running on 1 CPU of a Silicon Graphics Iris 4D/240GTX. The annealing program performs a single cooling cycle of simulated annealing and saves the low energy conformations it finds; for the same flavodoxin prediction this required 8 CPU h. Finally, the stored conformations are combined in a Boltzman-weighted average to produce the final prediction. This step took 2 min CPU time for flavodoxin.

3. Results

To test this method's accuracy for both exposed and buried residues, we used it to predict side-chain conformations for a set of nine proteins ranging in size from 45 to 323 residues (Tables 2 and 3; the flavodoxin prediction is illustrated in Fig. 1). In seven of these cases, all side-chains possessing χ torsions were moved and predicted simultaneously; for the largest two proteins (app and tln) only the most buried residues (roughly 1/3 of the residues in each protein) were included in the predictions. The predictions ranged in accuracy from 2·61 Å r.m.s. (worst) to 1·12 Å (best), with an overall r.m.s. error of 1·77 Å for the prediction set (see Table 2). As expected, the method worked considerably better for buried than exposed residues, predicting the core residues of app and tln with r.m.s. values of 1·12 Å and 1·28 Å, respectively. The overall r.m.s. error for core residues (see Methods for selection criteria) was 1·25 Å for the entire set of predictions. Apart from slight reductions in the tryptophan and phenylalanine r.m.s. the core predictions did not seem so much an improvement in the r.m.s. values of individual residues types, as an exclusion of the residue types for which the method performs badly (such as arginine and lysine; see Table 3).

Histograms of residue prediction errors measured as r.m.s., χ_1 and χ_2 deviations showed that the errors in the worst cases (pti and cro) were due primarily to a small fraction of the total residues predicted (Fig. 2). For pti (overall side-chain r.m.s. = 2·61 Å), 71% of the residues had r.m.s. values <2 Å, while 76% of the χ_1 predictions and 55% of the χ_2 predictions were within 40 degrees of correct. Similarly, in cro (r.m.s. = 2·39 Å) these fractions were 77%, 89% and 69%, respectively. The prediction errors giving rise to these proteins' high overall r.m.s. values were concentrated in only a few types of residues, and were mostly at the protein's surface (see Table 3). In pti, arginine (r.m.s. = 4·51 Å, number of cases $n = 6$), lysine (r.m.s. = 2·71 Å, $n = 4$) and glutamate (r.m.s. = 2·60 Å, $n = 2$) were the only residue types with r.m.s. values as large or larger than the overall value (2·61 Å), and were all located at the protein surface. Surface residues constituted the bulk of the remaining errors as well. The unusually high 2·17 Å r.m.s. for phenylalanine in pti (compared with 1·29 Å in the overall set) was due to a single surface phenylalanine which the program rotated out into solvent in its overzealous efforts to relieve van der Waals collisions. The prediction for this exposed residue had an r.m.s. error of 4·20 Å, while the other three phenylalanine residues in the protein were within 0·48 Å and 0·71 Å r.m.s. A similar error was made for a single surface tyrosine (r.m.s. = 3·43 Å, *versus* 0·19 Å to 0·94 Å r.m.s. for the other 3 tyrosine residues in pti). The only other residue types in pti with r.m.s. values greater than 1·5 Å were aspartate (r.m.s. = 1·81 Å, $n = 2$) and asparagine (r.m.s. = 1·56 Å, $n = 3$), all exposed.

The prediction errors responsible for the high overall r.m.s. value of cro (2·43 Å) followed the same pattern. Arginine (4·43 Å, $n = 5$), phenylalanine (3·35 Å, $n = 2$), and lysine (2·87 Å, $n = 6$) were the only residue types with r.m.s. values at or above the overall value, and in all these cases the errors were for surface residues largely unconstrained by packing. Of the two phenylalanine residues in cro, the one at the surface was predicted to be rotated into solvent (r.m.s. = 4·65 Å), while the one in the core was correct (r.m.s. = 0·92 Å). The remaining residue types with r.m.s. values above 1·5 Å were glutamine (2·20 Å, $n = 6$), tryptophan (2·10 Å, $n = 1$) and asparagine (1·61 Å, $n = 1$); all were at the surface. Exposed residues seemed to pose problems for the prediction method, not only because electrostatic interactions frequently were important for determining their conformations, but also because the program had no energy function representing the hydrophobic effect, and often mispredicted residues by failing to bury them appropriately. This was the case for the tryptophan in cro, which in the native structure lies flat in a shallow groove on the surface. The program predicted it in the correct orientation, but erroneously raised it slightly from the groove.

This clustering of errors in a small subset of surface residues prevailed over the entire set of predicted proteins. In the whole-protein predictions, the fraction of residues predicted within 2 Å r.m.s. ranged from 71 to 81%, while the fraction of χ_1 angles within 40 degrees was 57 to 89% (50 to 69% for χ_2). Once again, these ratios were markedly better in the buried-core predictions (app and tln): 91 to 92%, 81 to 82% and 74 to 81%, respectively, for r.m.s. χ_1 and χ_2. Lysine, arginine and glutamate were frequently the only residue types with errors larger than the overall value in each protein.

To assess the accuracy of predictions of each residue type, we calculated a "quality factor" for each residue type by dividing the expected r.m.s. error for random conformations by the actual error of the predictions (Fig. 3). The quality factor thus

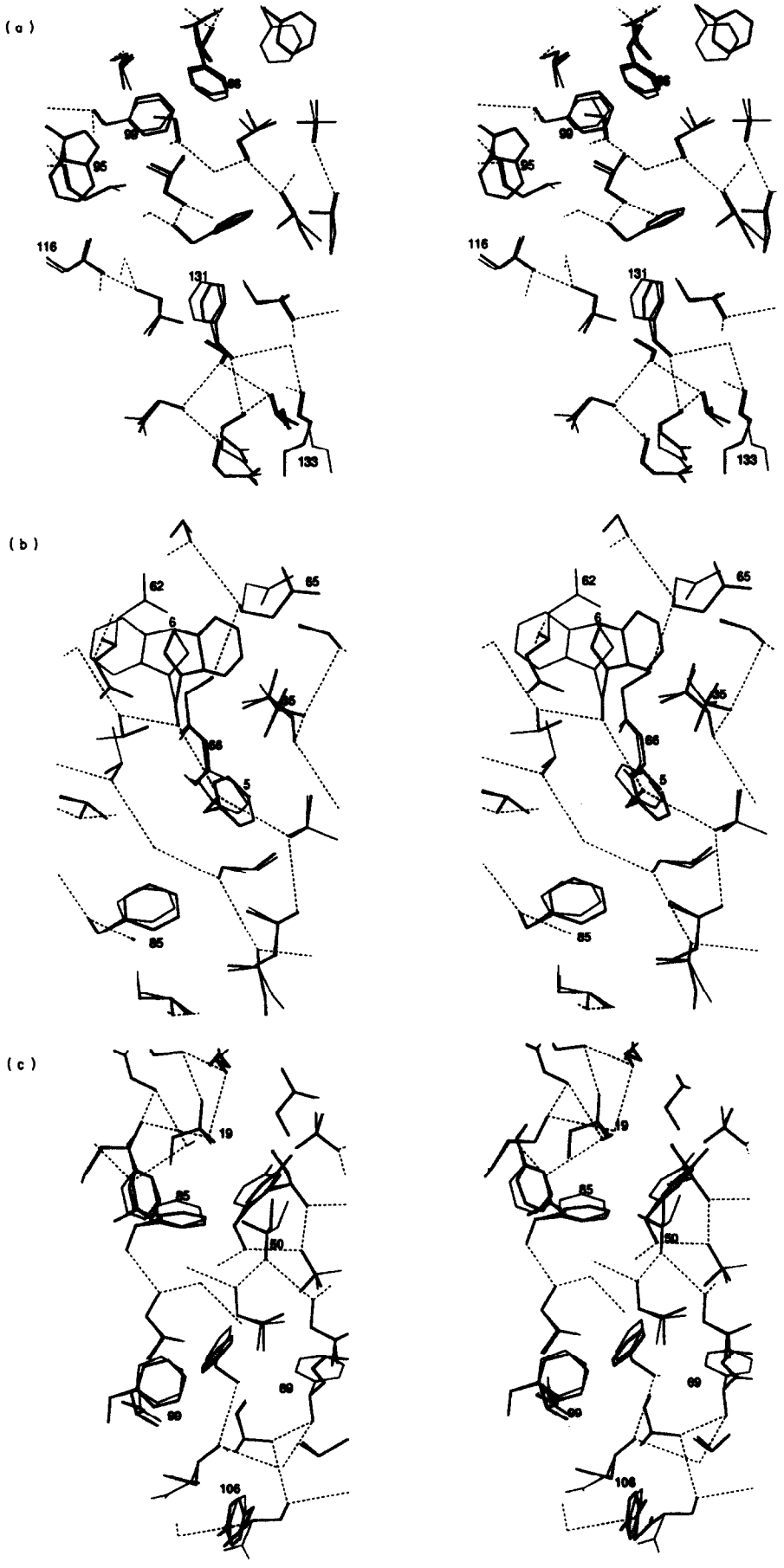
(a)
116
131
133
(b)
62
65
6
5
85
(c)
19
85
80
99
106

Table 2
r.m.s. deviations (Å) of predictions' side-chain atoms from the X-ray structures

Protein	Overall	Core residues
crn	1·65	1·23
pti	2·61	1·65
ctf	1·86	1·15
cro	2·39	1·15
rns	1·86	1·24
lz	1·62	1·08
fxn	1·90	1·53
tln	1·28	1·53
app	1·12	1·53
All cases	1·77	
Whole proteins only	1·97	
Core only		1·25

Whole proteins only refers to the set of 7 proteins for which all side-chains with free chi torsions were predicted simultaneously. In app and tln only the most buried residues (about 1/3 of each's total sequence) were predicted; to prevent the unpredicted residues from constraining the predictions, their side-chain atom co-ordinates were wholly deleted prior to simulated annealing, leaving empty space in their place.

indicates the ratio of improvement of the predictions' r.m.s. over random. Intriguingly, large hydrophobic side-chains were predicted best of all, especially tyrosine, phenylalanine, methionine, leucine and isoleucine. Charged residues, particularly lysine and arginine, and very small residues such as serine, threonine and cysteine had the lowest quality ratios. Poor prediction of serine and threonine residues resulted both from their small size (which gives them reduced steric hindrance) and the importance of hydrogen bonds for determining their conformation. Cysteine residues had a high error rate because the program made no effort to identify potential disulfide bridges, and effectively prohibited their formation by using the normal sulfur van der Waals radius for cysteine–cysteine interactions, which gives an energy minimum at an interatomic separation of 4·3 Å. Where disulfides did occur in native structures, the prediction typically placed one cysteine of the pair in the correct location, forcing the other into a rotamer 120° away from the correct position. Finally, comparison of residues' quality factor against side-chain volume reveals a loose correlation between prediction accuracy and the residues' size, especially for hydrophobic side-chains; charged and polar residues uniformly fell below this correlation line.

To resolve errors more clearly, we have prepared graphs of the r.m.s., χ_1 and χ_2 errors for each residue plotted against sequence position, for three proteins spanning the range of errors observed in the set: cro (2·39 Å r.m.s.), lz (1·62 Å) and app core (1·12 Å) (Fig. 4). The plot of χ_1 errors for cro shows the dominance of lysine and arginine in prediction errors, as well as the assignment of incorrect rotamers for a surface threonine (101) and the phenylalanine previously noted. The χ_2 plot is more interesting. Apart from some assignments of charged and polar residues to incorrect rotamers, two leucines appear to have been flipped in χ_2 (139° for Leu113; −152° for Leu120). Comparison of these conformations with the native reveals an interesting degeneracy in leucine's side-chain conformations: if χ_1 is rotated 30° to 40° and χ_2 turned 150° to 140°, a conformation is produced which superimposes the C^{δ} atoms nearly exactly on those of the initial structure, rendering it almost indistinguishable except for a slight shift of C^{γ}. In both Leu113 and Leu120 the predicted C^{γ} positions are within 1 Å of the native, while the C^{δ} superimpose with an r.m.s. value of 0·6 Å.

The by-now familiar theme of mispredicted lysine and arginine residues emerges once again from plots of r.m.s., χ_1 and χ_2 errors for lz. A single tryptophan of the four was mispredicted; sandwiched between a lysine and an arginine side-chain in the native structure, it was rotated into an incorrect rotamer in the solvent. Three of the six serine residues were predicted incorrectly: of these, two were totally exposed at the surface and one is involved in a hydrogen bond to a peptide N in the native structure. Two of the five isoleucine residues were mispredicted (89 and 106), one at the surface. The χ_2 plot shows a similar pattern of errors dominated by arginine and polar residues (particularly glutamine and asparagine). An additional isoleucine error is revealed: Ile23, located at the protein surface, has its $C^{\delta 1}$ rotated into the solvent.

Plots of errors in the prediction of the 100 most buried side-chains of app show the method's ability to sort out the complicated coupling of side-chain packing in the protein core, and to predict the conformations of the large side-chains that dominate it. The total r.m.s. error for the 16 phenylalanine side-chains in app's core is 0·91 Å; only one is mispredicted (Phe32). It is placed and oriented in approximately the right position, but is about 2 Å too close to Leu124. Any attempt at relieving bad contacts in the prediction (such as simple energy

Figure 1. Comparisons of flavodoxin prediction (bold lines) *versus* the X-ray structure (thin lines); the main-chain is shown as a C^{α} trace (broken lines). The prediction shown is the 1·90 Å r.m.s. flavodoxin whole-protein prediction, the worst of the set of 7 predictions generated for flavodoxin. This prediction had nearly the highest level of error for core residues (1·50 Å r.m.s.) in the entire set of test proteins. Core side-chains, especially large hydrophobic groups, were typically predicted within 1 Å of the correct position (e.g. Trp95, Phe99, Phe66, Tyr5, Phe85, Phe131, Ile116, Ile50), while surface side-chains, especially charged residues, were poorly predicted by packing energy optimization (e.g. Lys133, Glu62, Glu65; Trp6 was flipped 180° in χ_2 so that it pointed in the opposite direction in the surface groove it occupies). Small side-chains such as serine, valine and threonine appeared to be weakly constrained by packing forces, and were often predicted in incorrect rotamers (e.g. Val35).

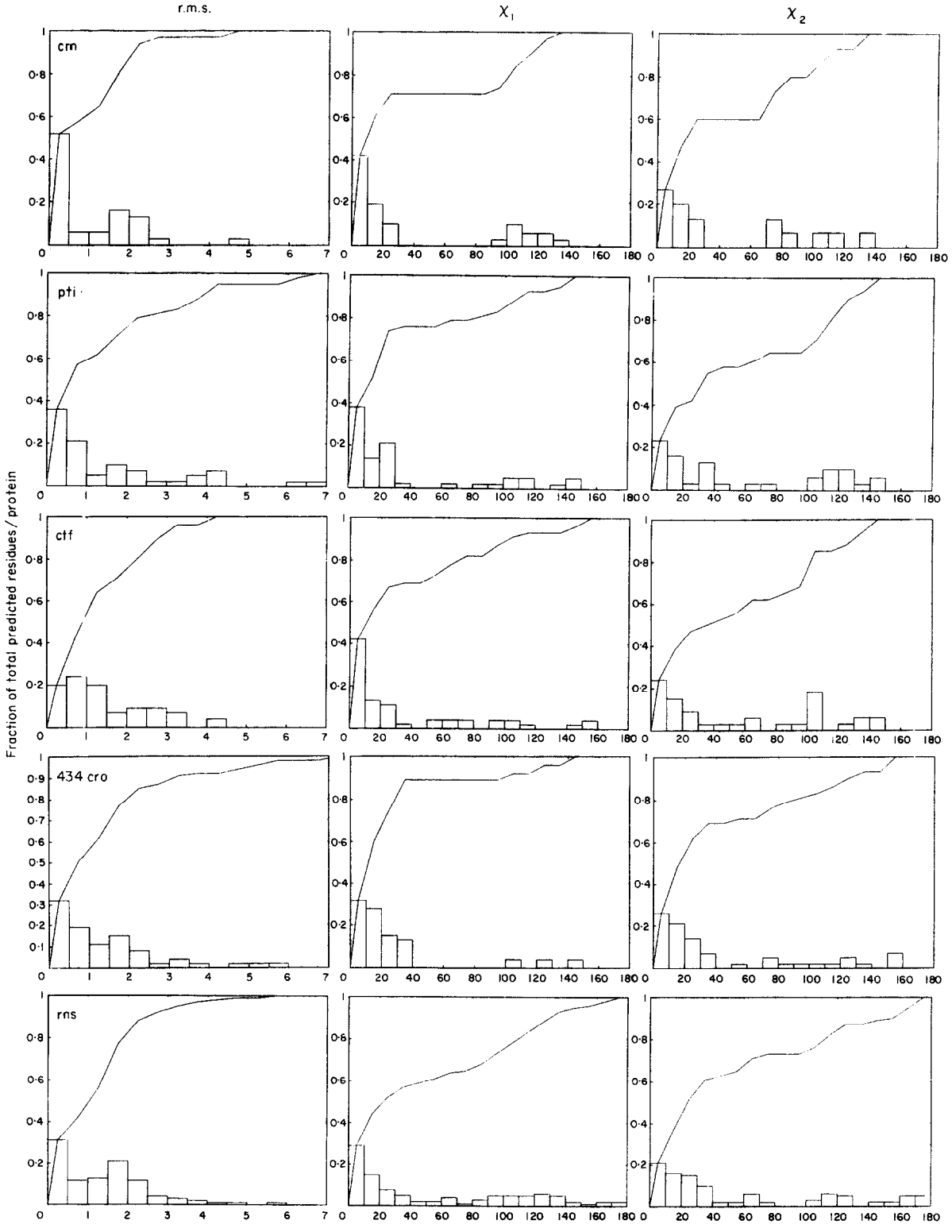
r.m.s.
χ_1
χ_2
cm
pti
ctf
434 cro
rns
Fraction of total predicted residues / protein

Fig. 2.

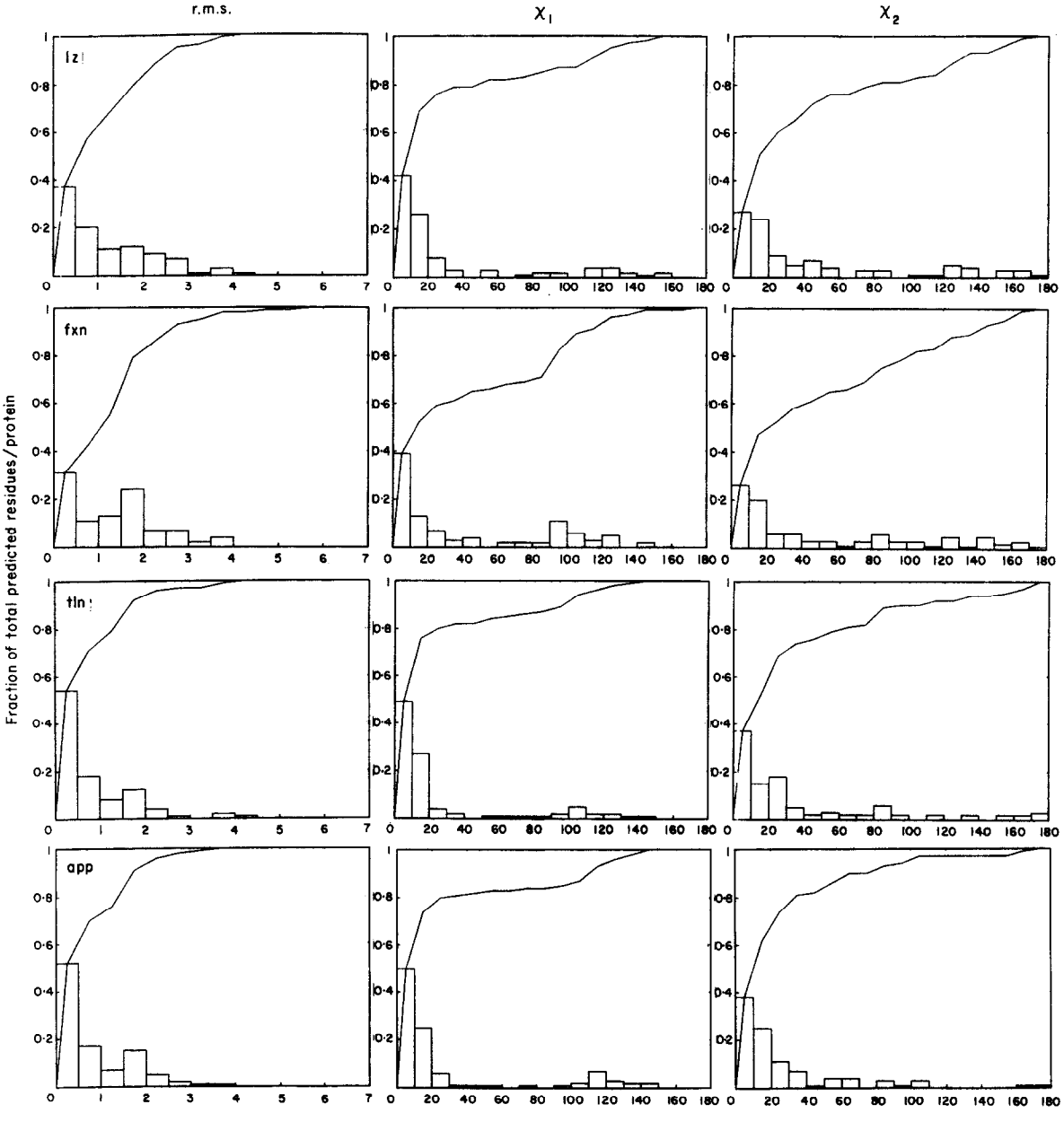

Figure 2. Histograms of prediction errors in each protein tested, measured as r.m.s. error/residue, χ_1 angle deviations from the X-ray structure, and χ_2 angle deviations from the X-ray structure. In each case the bar graph represents the distribution of predicted residues over the range of observed error magnitudes, while the line plot indicates the cumulative fraction of residues over the range of errors. In general, 60 to 90% of the predicted side-chains were within 2 Å r.m.s. of correct, and within 40° of the correct χ_1 angle. The 2 cases in which only core residues were predicted (app and tln) were significantly more accurate than the whole-protein predictions, reflecting the heightened importance of packing constraints in determining side-chain conformations in the core.

Table 3

r.m.s. deviations of predicted side-chains from the X-ray structures, subdivided by residue type (r.m.s. error/number of cases)

Amino acid	Protein									
	crn	pti	ctf	cro	rns	lz	fxn	tln	app	All
Trp				2·10/1		1·76/5	4·02/3			2·20/15
				2·10/1		0·53/4	2·82/2	0·82/3	0·91/3	1·41/13
Tyr	0·36/2	1·80/4		0·65/1	0·65/6	0·87/6	0·66/3			1·17/37
		2·43/2			0·82/3	0·66/3	0·67/2	1·77/7	0·89/8	1·33/25
Phe	0·85/1	2·17/4	1·03/1	3·35/2	0·46/3	0·85/2	0·66/5			1·29/41
	0·85/1	0·62/3			0·46/3	0·85/2	0·66/5	1·06/7	0·91/16	0·86/37
Met		0·30/1		1·23/3	1·59/4	0·67/2	1·49/5			1·27/17
				1·39/1	1·36/3	0·67/2	1·64/4	0·52/2		1·28/12
Ile	0·93/5	0·90/2	0·52/2	0·23/4	0·39/3	1·79/5	0·99/14			0·89/55
			0·52/2	0·24/3	0·39/3	1·79/5	1·05/12	0·67/13	0·26/7	0·90/45
Leu	0·31/1	0·44/2	1·50/7	1·10/8	0·26/2	0·39/8	1·21/8			1·00/63
			1·32/4	1·03/6	0·09/1	0·37/5	0·93/6	1·08/10	0·84/17	0·94/49
Val	0·30/2	0·05/1	0·89/8	0·19/3	1·25/8	0·76/8	1·56/10			0·97/74
			0·93/6	0·12/1	1·31/5	0·17/4	1·65/8	0·60/14	0·91/20	0·99/58
Cys	1·53/6	1·22/6		0·18/1	1·74/8	0·90/8	1·15/3			1·32/33
	1·29/4	1·69/3		0·18/1	1·74/8	1·02/6	1·15/3		0·23/1	1·38/26
Arg	3·48/2	4·51/6	1·04/1	4·43/5	3·83/4	2·38/14	1·72/2			3·40/35
				1·20/2		3·21/1		3·92/1		2·67/4
His					1·06/4	0·51/1				1·58/11
					1·08/2			2·01/5	1·55/1	1·76/8
Lys		2·71/4	2·87/9	2·87/6	2·90/10	1·81/5	2·89/10			2·72/46
							2·50/1	1·29/1	0·76/1	1·68/3
Gln		0·43/1		2·20/6	1·38/7	2·54/6	2·02/2			2·01/29
					0·71/1	0·93/1		0·43/1	2·16/6	1·81/9
Glu	2·12/1	2·60/2	1·84/9	1·20/4	1·98/5	0·91/3	1·83/19			1·73/48
				0·25/1		0·75/1	2·98/2	0·85/4	0·85/1	1·56/9
Asn	2·12/2	1·56/3	1·58/1	1·61/1	2·11/10	1·33/10	1·52/8			1·70/42
		1·55/2			2·45/2	1·62/1		1·62/3	1·72/4	1·81/12
Asp	0·96/1	1·81/2	0·59/4	0·93/1	1·21/5	0·78/8	1·80/9			1·31/43
					1·47/2	0·79/1		1·18/5	1·42/8	1·33/16
Thr	1·37/6	0·12/3	2·14/1	1·03/5	1·14/10	1·27/5	1·58/5			1·22/49
	1·49/2		2·14/1		1·56/4	0·24/2	0·07/1	0·65/8	1·56/6	1·24/24
Ser	0·13/2	0·38/1	1·21/2	0·31/2	1·27/15	1·19/6	1·14/8			1·17/55
					1·20/2	1·37/1	1·33/3	1·16/13	1·41/6	1·26/27

For all occurrences in each protein (upper line), and only those in the core (lower line).

minimization) would probably correct this error. The remaining χ_1 errors are due primarily to aspartate and glutamine (which were involved in internal hydrogen bonds in the native structure), and the small side-chains serine, threonine and valine. Three of the four mispredicted serine residues had hydrogen bonds in the native structure; the other was exposed at the surface. Of the four threonine residues placed in incorrect rotamers, three form hydrogen bonds to the main-chain in the native. Four of the 20 valine residues in app's core were mispredicted; this probably reflects valine's greater steric freedom, which makes it more difficult to predict by simple packing. The χ_2 errors are mainly flips: His54 is rotated $-161°$ from the correct χ_2, so that it lies in the same volume as in the native structure, but has lost the proper hydrogen bonding. Leu39 is flipped so that C^γ and one of the C^δ are closely aligned with the native, while the remaining C^δ is about 1 Å away from its native alternate. Leu121's C^δ atoms point out to the protein surface, and appear weakly constrained by packing. Leu122's error seems to have been caused by the shift in its neighbor Phe32 previously discussed. Finally, Leu284's prediction is a simple flip nearly indistinguishable from the native, of the kind described above for cro.

To assess the program's reliability and consistency, we have generated seven separate predictions for one protein (4fxn), each starting from different random conformations. These runs converged well, giving a consistent prediction regardless of starting conformation. All of these predictions had overall predicted side-chain r.m.s. values from the native structure of 1·59 to 1·90 Å, *versus* the random starting conformations' side-chain r.m.s. value of 3·00 to 3·34 Å. The internal r.m.s. value between different predictions ranged 1·15 to 1·55 Å, with very little deviation (0·2 to 0·8 Å) in the predictions of hydrophobic side-chains such as phenylalanine, tyrosine, isoleucine, leucine and valine. In contrast, charged residues had internal r.m.s. deviations around 2 Å. Starting from this observation that such residue types that were normally predicted poorly also had high internal r.m.s. deviations, we found that the internal r.m.s. deviation for each

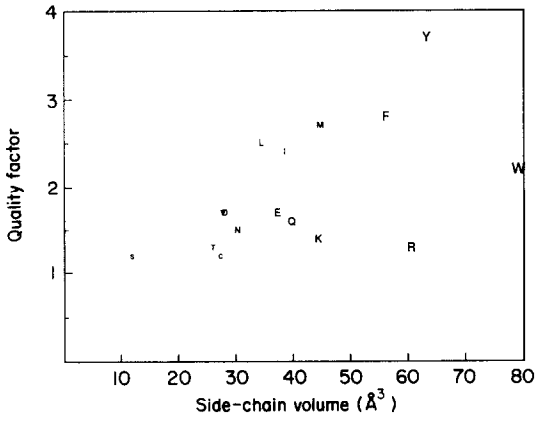

Figure 3. Prediction quality factors for each residue type, plotted against side-chain volume. The quality factor q for each residue type has been defined as the ratio of the r.m.s. for a random distribution of conformations divided by the overall r.m.s. of the predictions for that residue:

$$q = \text{r.m.s.}_{\text{random}}/\text{r.m.s.}_{\text{predictions}}.$$

Large hydrophobic side-chains were predicted most accurately, roughly proportionally to their size, while charged and polar side-chains were predicted less accurately. Cysteine residues were predicted less accurately than might be warranted by their size, due to the programs' systematic disregard for disulfide bridges, while tryptophan errors were often caused by the residue's frequent exposure at the protein surface.

residue over the set of predictions was a strong predictor of the accuracy of its predicted conformations against the native structure (Fig. 5). In general, a residue's r.m.s. from native was never significantly less than its internal r.m.s. deviation over the set of predictions from random starts. This observation provides a useful internal indication of probable errors that could offer a way to improved predictions. In general, the idea of using multiple predictions for detecting errors and improving accuracy appears to be a powerful tool for modelling (M. Levitt, personal communication).

4. Discussion

These results provide relatively clear measures of how strongly packing constraints determine the side-chain conformations of different residue types (see overall data, Table 3, or quality factor data, Fig. 3). Hydrophobic side-chains inside proteins, especially large residues such as phenylalanine and tyrosine, are generally constrained to unique conformations by packing forces. When located at the protein surface, however, such side-chains often have alternative conformations which in actual protein folding are ruled out by the hydrophobic effect. Small polar side-chains, such as serine and threonine, are typically constrained to one or two possible conformations, with hydrogen bonding selecting the appropriate one. Surface residues are poorly constrained by packing.

These results also indicate the importance of packing for determining the internal structure of proteins. Perhaps the most interesting result of this work is that there do not appear to be significant alternative ways of packing a given sequence of side-chains into the internal architecture established by the main-chain fold. Indeed, the native core structure appears to be a global minimum for packing energy, without significant alternatives. The observation that van der Waals interactions alone are sufficient to predict accurately and comprehensively the internal structure of proteins, in all of the cases examined so far, strongly suggests this conclusion. The convergence of the multiple flavodoxin predictions, from completely different starting points, to a single predicted core structure shows both that the program is successfully exploring the full conformational space, and that the energy function representing core packing interactions has a global minimum. The fact that this minimum matches the native structure confirms this view. This result is consistent with experimental and theoretical work indicating the importance of packing in determining core structure (Richards, 1977; Ponder & Richards, 1987; Lim & Sauer, 1989; Karpusas *et al.*, 1989; Gregoret & Cohen, 1990). Specifically, it suggests that packing constraints enforce a one-to-one correspondence between the main-chain fold and the pattern of side-chain packing that constitutes the protein core.

With this in mind, it may not be surprising that our method provides dramatically more accurate predictions of core side-chain conformation than existing methods, despite its total ignorance of electrostatics, the hydrophobic effect or statistical relationships (such as rotamers) commonly used to predict side-chains. It correctly predicts around 80% of core side-chains (within 1 Å r.m.s. of the correct position); most importantly, it correctly places nearly all large hydrophobic side-chains (Trp, Phe, Tyr, Ile, Leu) buried in the core. In sharp contrast with other methods, for which large side-chains are typically most problematic (Reid & Thornton, 1989), with 30 to 40% correctly predicted (2·41 Å r.m.s. error for flavodoxin side-chain atoms), most of our method's mistakes in protein cores are small side-chains, which have greater steric freedom and are often strongly determined by hydrogen bonding. Since it is the large side-chains which dominate the overall packing of the core, this accuracy bias towards larger residues is advantageous. Finally, the speed and simplicity of this method, which can be run from start to end in a day, allows multiple predictions to be generated using different starting points, as a means for identifying residues whose predictions are likely to be unreliable.

Although in this paper we have only presented data on prediction of side-chains in the context of correct main-chain co-ordinates, it is only logical to consider application of this method to homology modeling. The most obvious difficulty would be main-chain shifts which naturally occur when

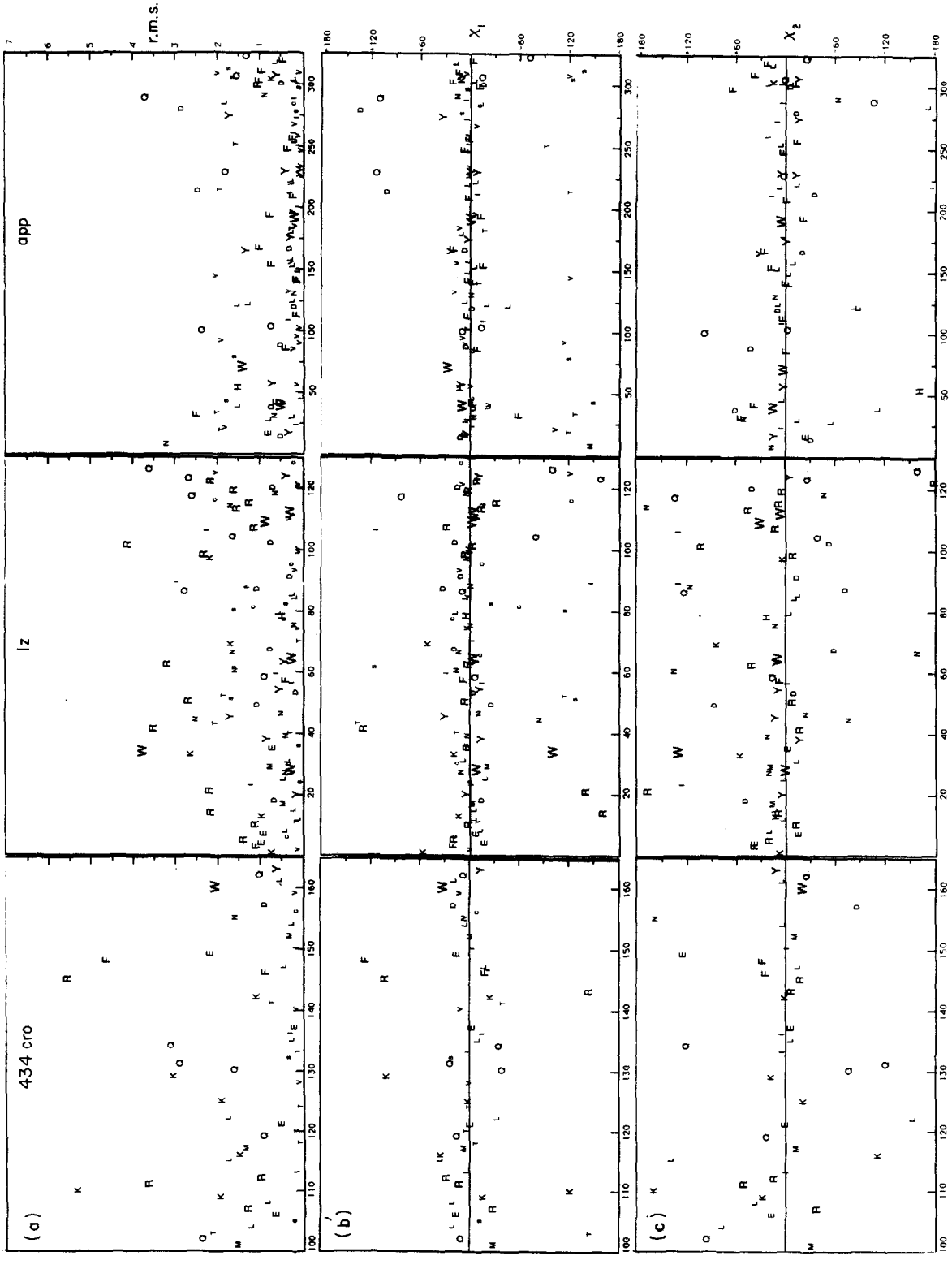

Figure 4. Prediction errors measured as (a) r.m.s. (in Å), (b) χ_1 deviations (in deg.), and (c) χ_2 deviations (in deg.), plotted against sequence position for 434 cro. lz and app. The letters representing the amino acids have been scaled so that their area is roughly proportional to their side-chain size.

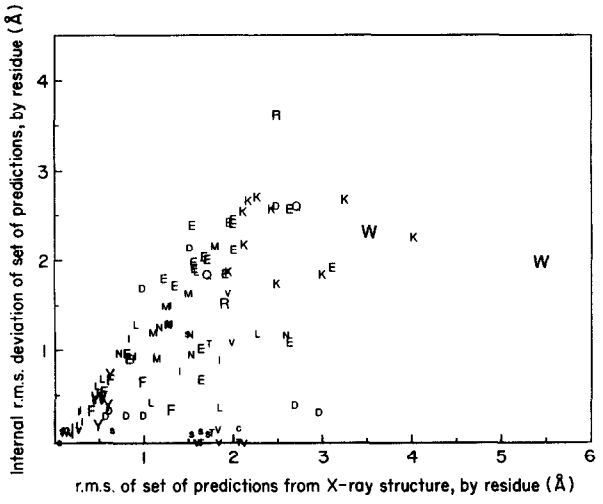

Figure 5. Internal deviations in multiple predictions of flavodoxin, against the overall deviation of these predictions from the X-ray structure. Seven predictions of flavodoxin were generated from different random starting points; here the internal pairwise deviation over the 7 predictions is plotted for each residue against its r.m.s. deviation from the X-ray structure. Nearly all of the large hydrophobic side-chains in the core were predicted in the same conformation in all predictions (internal r.m.s. <1 Å), and closely matched the native structure. Exposed side-chains, especially charged residues, had both much higher internal deviation within the prediction set and much higher error relative to the native structure.

mutations are introduced. Earlier theoretical studies have proposed that the general architecture of proteins should remain rigid, and that permissible mutations should preserve the volume of the protein interior (Ponder & Richards, 1987). However, recent experimental studies have shown that mutations changing the internal volume *are* permitted (Lim & Sauer, 1989), accommodated by shifts of not only side-chain but also main-chain atoms. Such main-chain shifts pose an acute problem for methods that use simple distance cut-off values to identify acceptable *versus* unacceptable packings (Ponder & Richards, 1987). Since this annealing method uses instead a sharply truncated 6 to 12 potential (7 kcal/mol maximum value per interaction pair) for assessing packings, and explores the full set of torsions for all side-chains simultaneously, it still seeks to identify the best possible packing even if this involves apparently "unacceptable" collisions with its starting main-chain model. The current version of the program allows side-chains to approach within 1 Å of main-chain atoms, to allow for potential shifts in main-chain co-ordinates of 1 to 2 Å from the starting model. Since proteins with a high degree of homology typically have main-chain differences for their core residues of less than 1 Å (Chothia & Lesk, 1986; Blundell *et al.*, 1987), this may provide sufficient flexibility for prediction in proteins produced by site-directed mutagenesis, and for proteins with significant homology. It is possible that successive cycles of side-chain prediction and main-chain relaxation (energy minimization with the side-chain torsions held fixed to their predicted values) could provide a method for both coping with and predicting main-chain shifts in homology modeling. We are currently testing such approaches for homologs with 20% or greater sequence identity, and examining the method's sensitivity to errors in the starting main-chain model.

One difficulty in the development and comparison of prediction methods is the lack of completely satisfactory measures of predictions' accuracy. Statistics on χ angle errors give disproportionate weight to smaller residues, since they include no information about side-chain size in the composite measurement. Thus, the critical issue of whether the big side-chains that dominate the core are correctly packed is simply absent from these measures. The χ error statistics for the flavodoxin prediction shown in Figure 1, for example, are comparable to those obtained for the prediction by Reid & Thornton (1989), yet direct examination of the core residues shows them to be dramatically different in accuracy (for comparison, see their Figs 4 and 5). r.m.s. deviations, on the other hand, are singularly slippery whenever data involving more than one residue type are being compared. An r.m.s. error of 1 Å has a dramatically different meaning for a serine than for a tyrosine residue; yet calculation of the overall r.m.s. mixes them in a single composite as if they were equivalent (although it does accord them appropriate weighting, in proportion to the number of atoms in each side-chain). As a result, the significance of a given r.m.s. value is different for different sequences. Furthermore, the relationship between r.m.s. value and human modelers' concept of "accuracy" is both highly compressed and non-linear. In the flavodoxin case, for example, completely random conformations have an average r.m.s. of 3·1 Å, while the r.m.s. value for our core prediction (with 80 to 90% of the side-chains correctly placed) is 1·5 Å. In general, how should one interpret a 1·9 Å r.m.s.? We wish to suggest that this question does not have a clear-cut answer, one must always probe further into the specifics of *what* was wrong, and what was right, usually by direct examination on graphics.

Finally, we wish to indicate some of the lines of work we are currently pursuing to extend and improve this method. The most obvious gap in the program is its total lack of provisions for prediction of surface residues. We are testing two different approaches. First, we are testing combination of this energy-based algorithm with statistical methods for side-chain prediction based on a library of 80 protein structures, which have proven successful for surface residue prediction (r.m.s. error of 1·76 Å for side-chain predictions on a similar set of proteins; M. Levitt, personal communication). Since conserved residues tend to preserve very similar conformations in homologous proteins (Lesk & Chothia, 1986), it may be possible to use the degree of conservation for each sequence position

(derived from multiple sequence alignment) to bias the random walk towards likely conformations, and thus improve the accuracy of the method. Second, we are adding electrostatics and the hydrophobic effect to our current program's potential function, for direct prediction of exposed residues within the main program. These additions are likely to improve its accuracy for buried residues as well, especially those involved in internal hydrogen bonds. In light of the observed prevalence of internal hydrogen bonds (Baker & Hubbard, 1984), inclusion of such interactions is essential for accurate prediction of these residues. A further problem experienced in the predictions reported here was the program's occasional attempts to relieve steric collisions by rotating hydrophobic side-chains completely into the solvent. If it is to be used for modeling surface residues, especially in homology modeling, the method must take the hydrophobic effect into account. In light of available algorithms for identifying potential disulfides (Pabo & Suchanek, 1986; Matsumura *et al.*, 1989), we are including recognition of disulfide bridges as well. Our main project, however, is incorporating main-chain movement explicitly in the prediction method, which may ultimately prove a more accurate basis for homology modeling than cyclic side-chain and main-chain optimization.

This work was done in Dr M. Levitt's group and supported by NIH grant GM-41455; we thank him for advice and constructive criticism. C.L. is a Howard Hughes Medical Institute predoctoral fellow. S.S. is supported by a Damon Runyon–Walter Winchell Cancer Research Fund fellowship, DRG-1019.

References

Artymiuk, P. J. & Blake, C. C. F. (1981). Refinement of Human Lysozyme at 1·5 Å Resolution. Analysis of Non-bonded and Hydrogen-bond Interactions. *J. Mol. Biol.* **152**, 737–762.

Atkins, P. W. (1986). *Physical Chemistry*, p. 587, Freeman, New York.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, 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.

Blow, D. (1983). Molecular Structure. Computer Cues to Combat Hypertension. *Nature (London)*, **304**, 213–214.

Blundell, T., Sibanda, B. L. & Pearl, L. (1983). Three-dimensional Structure, Specificity and Catalytic Mechanism of Renin. *Nature (London)*, **304**, 273–275.

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.

Borkatoti, N., Moss, D. S. & Palmer, R. A. (1982). Ribonuclease-A. Least-Squares Refinement of the Structure at 1·45 Å Resolution. *Acta Crystallogr. sect. B*, **38**, 2210–2217.

Bruccoleri, R. E. & Karplus, M. K. (1987). Prediction of the Folding of Short Polypeptide Segments by Uniform Conformational Sampling. *Biopolymers*, **26**, 137–168.

Brunger, A. T. (1988). Crystallographic Refinement by Simulated Annealing. Application to a 2·8 Å Resolution Structure of Aspartate Aminotransferase. *J. Mol. Biol.* **203**, 803–816.

Chothia, C. (1984). Principles that Determine the Structure of Proteins. *Annu. Rev. Biochem.* **53**, 537–572.

Chothia, C. & Lesk, A. M. (1986). The Relation Between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **5**, 823–826.

Chothia, C. & Lesk, A. M. (1987). Canonical Structures for the Hypervariable Regions of Immunoglobulins. *J. Mol. Biol.* **196**, 901–917.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Maruzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). The Predicted Structure of Immunoglobulin D1.3 and Its Comparison with the Crystal Structure. *Science*, **233**, 755–758.

Cohen, F. E. & Kuntz, I. D. (1987). Prediction of the Three-dimensional Structure of Human Growth Hormone. *Prot. Struct. Funct. Genet.* **2**, 162–166.

Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L. & Smith, K. A. (1986). Structure-activity Studies of Interleukin-2. *Science*, **234**, 349–352.

Connolly, M. (1985). Computation of Molecular Volume. *J. Amer. Chem. Soc.* **107**, 1118–1124.

Delbaere L. T., Brayer, G. D. & James, M. N. (1979). Comparison of the Predicted Model of α-lytic Protease with the X-ray Structure. *Nature (London)*, **279**, 165–168.

Gelin, B. R. & Karplus, M. K. (1975). Side-chain Torsional Potentials and Motion of Amino Acids in Proteins: Bovine Pancreatic Trypsin Inhibitor. *Proc. Nat. Acad. Sci., U.S.A.* **72**, 2002–2006.

Gelin, B. R. & Karplus, M. K. (1979). Side-chain Torsional Potentials: Effect of Dipeptide, Protein, and Solvent Environment. *Biochemistry*, **18**, 1256–1268.

Greer, J. (1981). Comparative Model-building of the Mammalian Serine Proteases. *J. Mol. Biol.* **153**, 1027–1042.

Greer, J. (1985) Model Structure for the Inflammatory Protein C5a. *Science*, **228**, 1055–1060.

Gregoret, L. M. & Cohen, F. E. (1990). Novel Method for the Rapid Evaluation of Packing in Protein Structures. *J. Mol. Biol.* **211**, 959–974.

Hendrickson, W. A. & Teeter, M. M. (1981). Structure of the Hydrophobic Protein Crambin Determined Directly from the Anomolous Scattering of Sulfur. *Nature (London)*, **290**, 107–113.

Holmes, M. A. & Matthews, B. W. (1982). Structure of Thermolysin Refined at 1·6 Å Resolution. *J. Mol. Biol.* **160**, 623–639.

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.

Jones, T. A. & Thirup, S. (1986). Using Known

Substructures in Protein Model Building and Crystallography. *EMBO J.* **5**, 819–822.

Karpusas, M., Baase, W. A., Matsumura, M. & Matthews, B. W. (1989). Hydrophobic Packing in T4 Lysozyme Probed by Cavity-filling Mutants. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8237–8241.

Kirkpatrick, S., Gelatt, C. D., Jr & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science,* **220**, 671–680.

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal Domain of the Ribosomal Protein L7/L12 from Escherichia coli at 1·7 Å. *J. Mol. Biol.* **195**, 555–579.

Lesk, A. M. & Chothia, C. (1986). *Phil. Trans. Roy. Soc. ser. A,* **317**, 345–356.

Levitt, M. (1983*a*). Molecular Dynamics of Native Protein: Computer Simulation of Trajectories. *J. Mol. Biol.* **168**, 595–620.

Levitt, M. (1983*b*). Protein Folding by Constrained Energy Minimization and Molecular Dynamics. *J. Mol. Biol.* **170**, 723–764.

Lim, W. A. & Sauer, R. T. (1989). Alternative Packing Arrangements in the Hydrophobic Core of Lambda Repressor. *Nature (London),* **339**, 31–36.

Matsumura, M., Becktel, W. J., Levitt, M. & Matthews, B. W. (1989). Stabilization of Phage T4 Lysozyme by Engineered Disulfide Bonds. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 6562–6566.

McCormick, F., Clark, B. F. C., la Cour, T. F. M., Kjedgaard, M., Norskov-Lauritsen, L. & Nyborg, J. (1985). A Model for the Tertiary Structure of p21, the Product of the *ras* Oncogene. *Science,* **230**, 78–82.

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.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1089.

Mondragon, A., Wolberger, C. & Harrison, S. C. (1989). Structure of phage 434 Cro Protein at 2·35 Å Resolution. *J. Mol. Biol.* **205**, 179–188.

Narayana, S. V. & Argos, P. (1984). Residue Contacts in Protein Structures and Implications for Protein Folding. *Int. J. Pept. Prot. Res.* **24**, 25–39.

Novotny, J., Bruccoleri, R. & Karplus, M. K. (1984). An Analysis of Incorrectly Folded Protein Models: Implications for Structural Predictions. *J. Mol. Biol.* **177**, 787–818.

Pabo, C. O. & Suchanek, E. G. (1986). Computer-aided Model-building Strategies for Protein Design. *Biochemistry,* **25**, 5987–5991.

Pearl, L. H. & Taylor, W. R. (1987). A Structural Model for the Retroviral Proteases. *Nature (London),* **329**, 351–354.

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.

Press, W. H., Flannery, B. R., Teukolsky, S. A. & Vetterling, W. T. (1986). Combinatorial Minimization: Method of Simulated Annealing. In *Numerical Recipes,* pp. 326–334, Cambridge University Press, Cambridge.

Read, R. J., Brayer, G. D., Jurasek, L. & James, M. N. G. (1984). Critical Evaluation of Comparative Model Building of Streptomyces Griseus Trypsin. *Biochemistry,* **23**, 6570–6575.

Reid, L. & Thornton, J. M. (1989). Rebuilding Flavodoxin from Cα Coordinates: A Test Study. *Proteins,* **5**, 170–182.

Richards, F. M. (1977). Areas, Volumes, Packing and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Sibanda, B. L. & Thornton, J. M. (1985). Beta-hairpin Families in Globular Proteins. *Nature (London),* **316**, 170–172.

Singh, J. & Thornton, J. M. (1990). SIRIUS. An Automated Method for the Analysis of the Preferred Packing Arrangements Between Protein Groups. *J. Mol. Biol.* **211**, 595–615.

Smith, W. W., Burnett, R. M., Darling, G. D. & Ludwig, M. L. (1977). Structure of the Semiquinone Form of Flavodoxin from Clostridium MP. Extension of 1·8 Å Resolution and Some Comparisons with the Oxidized State. *J. Mol. Biol.* **17**, 195–225.

Strynadka, N. C. & James, M. N. G. (1988). Two Trifluoperazine-binding Sites on Calmodulin Predicted from Comparative Molecular Modeling with Troponin-C. *Prot. Struct. Funct. Genet.* **3**, 1–17.

Subbiah, S. & Harrison, S. C. (1989). A Simulated Annealing Approach to the Search Problem of Protein Crystallography. *Acta Crystallogr. sect. A,* **45**, 337–342.

Summers, N. L. & Karplus, M. K. (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. K. (1987). An Analysis of Side-chain Orientations in Homologous Proteins. *J. Mol. Biol.* **196**, 175–198.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987*a*). Knowledge Based Modelling of Homologous Proteins, Part I: Three-dimensional Frameworks Derived from the Simultaneous Superposition of Multiple Structures. *Prot. Eng.* **1**, 377–384.

Sutcliffe, M. J., Hayes, F. R. R. & Blundell, T. L. (1987*b*). Knowledge Based Modelling of Homologous Proteins, Part II: Rules for the Conformations of Substituted Side-chains. *Prot. Eng.* **1**, 385–392.

Tang, J., James, M. N. G., Hsu, I. N., Jenkins, J. A. & Blundell, T. L. (1978). Structural Evidence for Gene Duplication in the Evolution of the Acid Proteases. *Nature (London),* **271**, 618–621.

Taylor, W. R. (1988). Pattern Matching Methods in Protein Sequence Comparison and Structure Prediction. *Prot. Eng.* **2**, 77–86.

Van Hemmen, J. L. & Morgenstern, I. (1983). Editors of Heidelberg Colloquium on Spin Glasses. *Lecture Notes in Physics,* vol. 192, Springer-Verlag, Berlin.

Van Kampen, N. G. (1981). In *Stochastic Processes in Physics and Chemistry,* Elsevier, Amsterdam.

Warme, P. K. & Morgan, R. S. (1978*a*). A Survey of Atomic Interactions in 21 Proteins. *J. Mol. Biol.* **118**, 273–287.

Warme, P. K. & Morgan, R. S. (1978*b*). A Survey of Amino Acid Side-chain Interactions in 21 Proteins. *J. Mol. Biol.* **118**, 289–304.

Weber, I. T., Miller, M., Jaskolski, M. Leis, J., Skalka, A. M. & Wlodawer, A. (1989*a*). Molecular Modelling of the HIV-1 Protease and its Substrate Binding Site. *Science,* **243**, 928–931.

Weber, I. T., Shabb, J. B. & Corbin, J. D. (1989*b*). Predicted Structures of the cGMP-dependent Protein Kinase: A Key Alanine/threonine Difference in Evolutionary Divergence of cAMP and cGMP Binding Sites. *Biochemistry,* **28**, 6122–6127.

Wlodawer, A., Walter, J., Huber, R. & Sjölin, L. (1984). Structure of Bovine Pancreatic Trypsin Inhibitor. Results of Joint Neutron and X-ray Refinement of Crystal Form II. *J. Mol. Biol.* **180**, 301–329.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by A. Klug