Accurate Modeling of Protein Conformation by Automatic Segment Matching

Michael Levitt

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

(Received 22 August 1991; accepted 9 March 1992)

Segment match modeling uses a data base of highly refined known protein X-ray structures to build an unknown target structure from its amino acid sequence and the atomic co-ordinates of a few of its atoms (generally only the C^α atoms). The target structure is first broken into a set of short segments. The data base is then searched for matching segments, which are fitted onto the framework of the target structure. Three criteria are used for choosing a matching data base segment: amino acid sequence similarity, conformational similarity (atomic co-ordinates), and compatibility with the target structure (van der Waals' interactions).

The new method works surprisingly well: for eight test proteins ranging in size from 46 to 323 residues, the all-atom root-mean-square deviation of the modeled structures is between 0·93 Å and 1·73 Å (the average is 1·26 Å). Deviations of this magnitude are comparable with those found for protein co-ordinates before and after refinement against X-ray data or for co-ordinates of the same protein in different crystal packings. These results are insensitive to errors in the C^α positions or to missing C^α atoms: accurate models can be built with C^α errors of up to 1 Å or by using only half the C^α atoms.

The fit to the X-ray structures is improved significantly by building several independent models based on different random choices and then averaging co-ordinates; this novel concept has general implications for other modeling tasks. The segment match modeling method is fully automatic, yields a complete set of atomic co-ordinates without any human intervention and is efficient (14 s/residue on the Silicon Graphics 4D/25 Personal Iris workstation).

Keywords: modeling conformation; data base; model building; side-chain positions

1. Introduction

Knowledge of protein conformation facilitates understanding of protein function. Specific interaction between a protein molecule and small molecules or other protein molecules depends on complementary fitting of surfaces in three dimensions (for example, the interaction of enzyme and inhibitor, receptor and hormone, or antibody and antigen). Determining structure by crystallography or nuclear magnetic resonance spectroscopy is a time-consuming procedure and, at present, there are many more known amino acid sequences than known protein structures; determination of sequence is always likely to be easier than determination of structure. Fortunately, proteins with different sequences have similar conformations provided the sequences are homologous. Thus, a general method for building the side-chains of an unknown protein onto the main-chain framework of another related protein would have widespread application.

Immediately after the first two unrelated protein structures were determined by X-ray crystallography, myoglobin by Kendrew *et al.* (1960) and lysozyme by Phillips *et al.* (1966), it became clear that the same local or secondary structure motifs recur in very different tertiary structures. The two local structures identified then, α-helices and β-sheets, have since been found to occur in all known globular protein conformations. The next type of local structure to be recognized was the reverse turn (Venkatachalam, 1968). With these three types of

0022-2836/92/140507-27 $03.00/0

secondary structure, the main-chain conformations of about three-quarters of all protein residues have a recognizable, standard conformation. Early energy calculations on side-chain conformations (Gelin & Karplus, 1975) demonstrated that side-chains adopt low-energy conformations close to those expected for the free amino acid. Subsequent analysis (Janin *et al.*, 1978) showed that there were relatively few common side-chain conformations; more recent work on a larger sample of very well-refined proteins confirmed that the commonly occurring conformations are very close to standard low-energy rotamers (Ponder & Richards, 1987).

While these studies showed that proteins generally use a common "vocabulary" of main and side-chain conformations, it was not clear how to take advantage of the ever increasing number of protein three-dimensional structures being solved by X-ray crystallography. The breakthrough in the use of this information was provided by Jones & Thirup (1986), who added a search for conformations of segments of known protein into Jones' program FRODO (Jones, 1978) used to fit a model polypeptide chain to electron density. Their idea was to use pre-existing segments of protein structure as an aid to fitting an atomic model of a new protein to its electron density map. After initial tests, they concluded that recurring conformations are not limited to the recognized regions of secondary structure: it is possible to find almost all segments of local main-chain conformation in other protein structures. Use of segments as an aid to molecular modeling was quickly implemented by others and used to fit both electron density maps (Finzel *et al.*, 1989*b*,) and model proteins from partial information (Blundell *et al.*, 1987). In all these schemes, the data base of known segments is used in conjunction with manual manipulation of the structure through interactive molecular graphics.

Impressed by the general power of these methods, I chose to implement a completely automatic segment modeling method, called segment match modeling or SMM†. The advantages of such an automatic method are as follows. (1) It can be tested on a large number of known protein structures provided that all information on the protein under consideration is removed from the data base; such objective modeling of known proteins is impossible with manual or semi-automatic modeling as anyone with the necessary expertise may be familiar with the protein being modeled. (2) It can be used repeatedly to generate a family of plausible models; manual modeling often leads to a favorite model to the exclusion of all other possible models. (3) It can be used properly by a novice, not requiring any expertise on the part of the user.

Segment match modeling is part of a general scheme for modeling protein conformation referred to as "black-box" modeling. Imagine a computer program (the "black box") that reads in the amino acid sequence of a protein together with an incomplete set of atomic co-ordinates and then outputs the positions of all the missing atoms. The success of this program can be judged by the degree of similarity of the calculated co-ordinates to the known atomic co-ordinates of the particular molecule. The difficulty of accomplishing the modeling task obviously depends on the number of missing atoms. At one end of the scale, it should be trivial to calculate the co-ordinates for a single missing atom. At the other end of the scale, producing all atomic co-ordinates from the amino acid sequence alone would constitute a solution to the protein folding problem. Thus, this modeling scheme allows a continuous transition from an easy, soluble problem through more challenging modeling to actual prediction of protein conformation from amino acid sequence.

Here, we focus on a task of intermediate difficulty and attempt to derive the co-ordinates for all the atoms from the amino acid sequence and the co-ordinates of all or some of the α-carbon atoms. Starting with these co-ordinates ensures that the overall chain path is correct. The conformation of the main-chain, in particular the orientation of the peptide groups, and the conformations of the side-chains remain to be determined by the program.

Segment match modeling follows the work of Jones & Thirup (1986) and builds a model for the target structure from a data base of known structures. To do this, the target structure is broken into short segments of sequence. These segments, which must have atomic co-ordinates determining their shape, are used to select segments of matching shape in the data base. Because each segment in the data base comes from a known protein, the conformation of all atoms in the segment is known. These segment co-ordinates are fitted into the growing target structure. The process is repeated until all atomic co-ordinates of the target structure are obtained. During the match modeling procedure care is taken to exclude all information about the target protein from the data base. This information is used subsequently to assess the accuracy of the modeling by comparing the calculated and known atomic co-ordinates.

A number of decisions have to be made during this construction process (where to start, which known segments are the best matches, etc.). These decisions are dealt with by a simple and surprisingly powerful general concept: make choices at random and average. More specifically, segment match modeling starts at a random position in the sequence, continues at random and each time randomly chooses one of the good segment matches. The procedure is repeated to give a number of independently built models that are then averaged to give a mean model for the target structure. Such averaging also helps highlight the significant features that occur repeatedly in all models; for these atoms, the variance from the mean co-ordi-

†Abbreviations used: SMM, segment match modeling; r.m.s., root-mean-square.

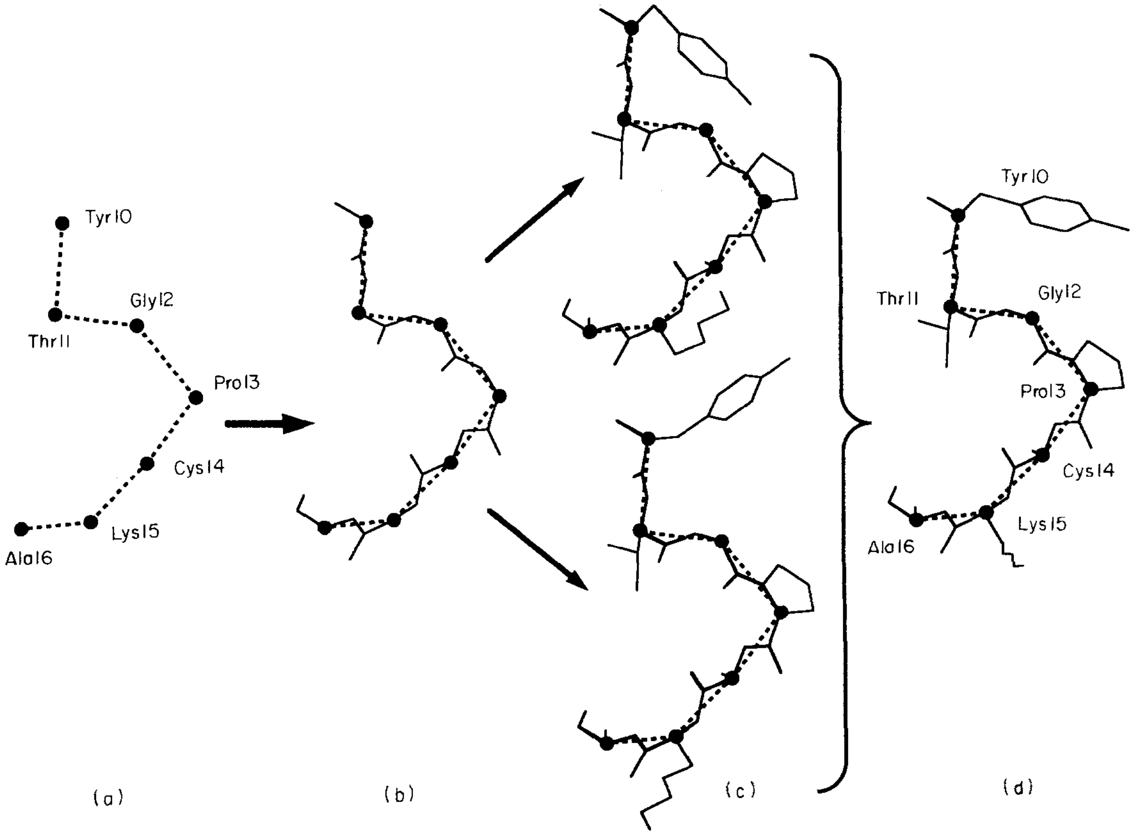


Figure 1. An illustration of the segment match modeling algorithm that is used to model protein conformations from C^α co-ordinates. (a) The rough conformation of a segment of polypeptide chain is defined by the positions of the C^α atoms shown as filled circles connected by broken virtual bonds. In the example, it is assumed that the positions of all C^α atoms are known. (b) In the 1st phase of modeling, the main-chain atoms are added to the C^α framework. (c) In the 2nd phase the side-chain atoms are added to the completed main-chain. In both phases the residues to be modeled are chosen in random order. The process is repeated 10 times to obtain 10 independent models. (d) The atomic co-ordinates of these models, which are different due to the random order of building and statistical choice of segments, are then averaged to give a mean model. The final model is obtained by limited energy minimization, which enforces good stereochemistry.

nates will be small and this indicates greater reproducibility.

Application of segment match modeling to eight different proteins gives modeled target structures with root-mean-square deviations (r.m.s.) of between 0·93 Å and 1·73 Å (1 Å = 0·1 nm) from the corresponding known structure (for all atoms). This deviation is small and many of the side-chains are seen to be correctly predicted in superimposed stereo views of the modeled and known structures. These results, which are obtained by a completely automatic method, are not sensitive to errors in the C^α positions or to missing C^α atoms: good models can be built with C^α errors of up to 1 Å or by using only half the C^α atoms.

2. Methods

The segment match modeling method is simple consisting of the following stages: make the data base, build several independent models of the target structure by matching short segments of polypeptide chain, average these models, refine the mean co-ordinates and analyze the fit of the final model to the X-ray structure. The first 3 stages are done by the program SegMod, using the procedure described below and illustrated by Fig. 1; the last 2 stages are done by the energy minimization program, ENCAD (Levitt, 1983*a*). In this work, segment means part of a protein consisting of consecutive, connected residues. The protein being modeled is referred to as the target. The target segment is the segment being modeled; the growing target structure is the set of atomic positions that is known or has been modeled at the current stage of the procedure.

(a) *The data base*

The data base of known proteins used is given in Table 1. It contains a total of 76 proteins, comprising 12,288 amino acid residues and 93,569 atoms. The proteins selected have all been solved to a resolution of 2 Å or better. Duplicates and minor modifications of the same protein have been excluded, but more than one member of a family of homologous proteins is allowed. Thus, only 1 of the more than 10 refined trypsin structures is included, but the co-ordinates for each of the homologous proteins, chymotrypsin and elastase, are included. The co-ordinates are all taken from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977); references to the individual structures are given in Table 1.

Because the atomic co-ordinates in the data base are to provide co-ordinates for segments of the target protein, it is important that the co-ordinates of all atoms be present and that there be no duplicated atoms. Often, sets of highly refined protein co-ordinates have both missing atoms and duplicated atoms (the former, because not all atoms are resolved even at the highest resolution, and the latter because side-chains can have multiple conformations seen at high resolution). After reading in the atomic co-ordinates from highly compressed co-ordinate files (5 bytes/atom; unpublished results), they are sorted into a standard atom order to detect efficiently any missing or duplicated atoms. Positions of chain gaps are also noted to prevent the selection of segments that span chain boundaries. Distances are calculated between all pairs of C^α atoms separated by from 2 to 19 residues; this distance information is used in the initial matching of segment shapes (see below). Co-ordinate input, sorting, checking and distance calculations are sufficiently rapid to be done every time (total time of 10 s on the Silicon Graphics 4D/25). This allows the list of proteins that are to be used for the data base to be changed at will. When testing the method on target proteins whose 3-dimensional structure is known, all information on the target protein is omitted from the data base.

(b) *Building target models*

Modeling of a particular target protein requires the amino acid sequence and a partial set of known atomic co-ordinates. For test runs, the known co-ordinates were taken as all the C^α atoms or a subset of these atoms. For general modeling, any partial set of co-ordinates is valid input. Segment match modeling is divided into 2 stages: first model any missing main-chain atoms (N, H, C^α, C, O), and then model any missing side-chain atoms. The protocol used for each of these stages is similar and consisted of the following 5 steps. (1) Choose the segment of the target protein to be modeled. (2) Find plausible matches to this segment in the data base of known proteins. (3) Sort the list of matches using as criteria the r.m.s. deviation from the target segment and the van der Waals' interaction energy with parts of the target protein already built. (4) Select a best match from the list of good matches. (5) Copy co-ordinates of some of the atoms from the selected segment to the growing target protein structure. This process is illustrated by Fig. 2.

(i) *Choose the segment of the target protein to be modeled*

The segment to be modeled is defined by the central residue, i, and the length of the segment, L. The central residue is chosen at random from those residues having at least 1 missing atom (either main-chain or side-chain, depending on the stage of the procedure). The start and end points of the segments are then chosen to ensure that there are sufficient constraints to guide the matching process. This is done by specifying the required number of known residues in the segment (L), and then finding a total of L consecutive C^α atoms that already have co-ordinates (either given as input or built by modeling) and are closest along the sequence to the central residue and in the same polypeptide chain. This requirement is most severe when there are many missing C^α atoms. For example, consider $L = 5$ and provision of known co-ordinates for every third C^α atom (say, residues 1, 4, 7, 10, 14, 17, 20, etc.). To model the main-chain of residue 9, use is made of the known co-ordinates of residues 4, 7, 10, 14 and 17, giving a target segment extending from residue 4 to 17. If the residue to be modeled is near the beginning or end of a polypeptide chain, the L residues chosen will not necessarily bracket the missing residue. Note, that as the modeling proceeds, the positions of C^α atoms located by the procedure can subsequently be used for further modeling. Thus, to model the side-chain of this same residue 9, the main-chain atoms of all residues will have been located and the 5 C^α atoms used for the template will be those of residues 7, 8, 9, 10 and 11. The optimum value of L depends on the stage of the modeling and provision is made to use different L values for main-chain and side-chain atoms. When 3 or more consecutive C^α atoms are missing, the procedure is modified to ensure that the central residue is adjacent to at least 1 located C^α atom. This is done by postponing the modeling of residues with 2 missing C^α neighbors until one or both of the neighboring C^α atoms have been located. Other schemes are possible within this framework, for example, model the side-chains of certain residue types first, model interior residues first, etc. Tests showed that the best results are obtained when selecting the central residue completely at random.

(ii) *Plausible segment matches to the data base*

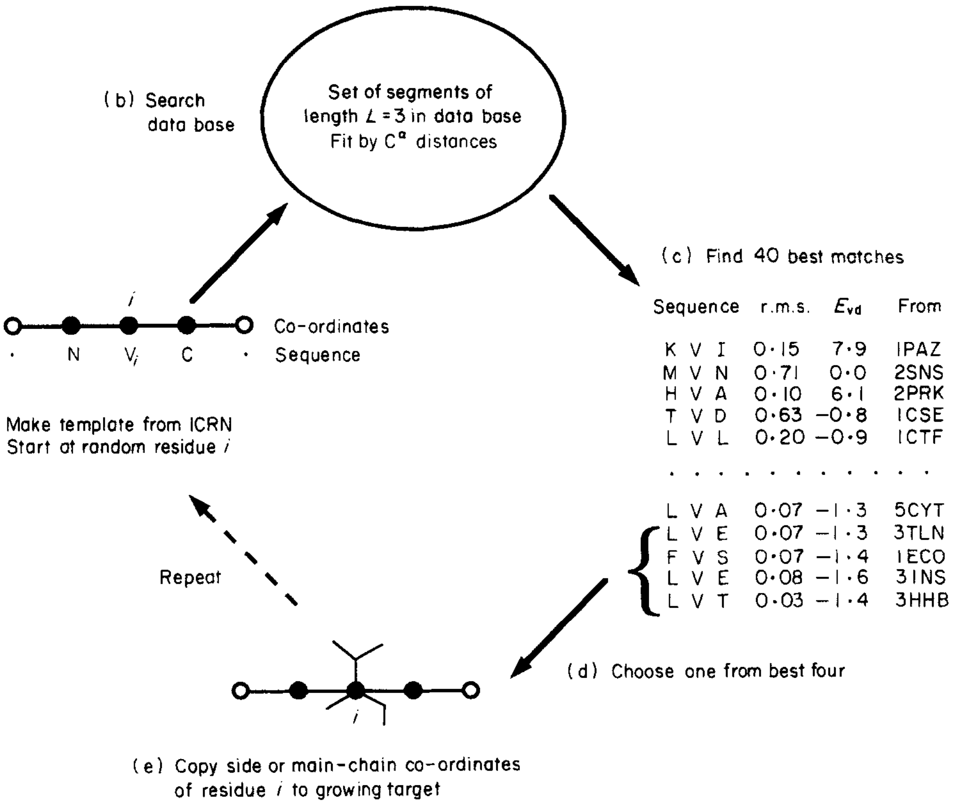
Segments that match the target segment are found by searching the entire data base and selecting matches on the basis of both the amino acid sequence and the conformation of the C^α-chain. Matching sequence is found by a template, which restricts the allowed amino acid at each position of the segment to be a specific amino acid or to belong to a particular class of amino acids. For example, the template "....." permits each of the residues in a segment of $L = 5$ to be any amino acid (denoted by the "." symbol), while the template "..A.." restricts the central amino acid to be alanine. Complicated templates can be specified and used to find matches. Here just 2 templates are used: when building the main-chain there are no restrictions (template is "....."); and when building a side-chain, the residue being modeled (usually the central residue of the segment) must be the same as in the sequence of the target protein (template was "..X..", where the central residue is "X").

Segments of matching conformation are selected by the method of Jones & Thirup (1986) in which inter-C^α distances are matched in preference to C^α positions. Matching distances has the advantage of being independent of segment orientation. A segment of 5 C^α atoms defines only 6 distances between the C^α atoms separated by 2 or more residues (1...3, 2...4, 3...5, 1...4, 2...5, 1...5) and these can be matched very quickly (for a segment of length L, there are $(L-1)(L-2)/2$ such distances to match). Inter-C^α distance matching does have 2 deficiencies: a structure and its mirror image will have identical distances and match well but may have very different conformations; and the shape of the polypeptide chain is determined by more than the C^α positions.

(iii) *Root-mean-square deviation and packing criteria*

The rapid search for matching segments using inter-C^α distances is augmented by 2 additional criteria: the co-ordinate deviation of atomic positions is calculated and the van der Waals' interactions are scanned to assess packing. These tests, which are time-consuming, are applied to only the 40 data base segments found to have the lowest inter-C^α distance deviations in the search of the entire data base. The co-ordinate deviation (as opposed to the C^α distance deviation) between data base and target segment is calculated for each of the 40 data base segments using the very efficient and stable best-fit algorithm due to Kabsch (1978).

The atoms used in the r.m.s. deviation calculation depend on the stage of modeling: when building main chain, the known C^α positions are used (known by initial


(b) Search data base
Set of segments of length L = 3 in data base Fit by C^{α} distances
(c) Find 40 best matches
i
Co-ordinates
· N V_i C · Sequence
(a) Make template from ICRN Start at random residue i
Repeat
(d) Choose one from best four
i
(e) Copy side or main-chain co-ordinates of residue i to growing target

Sequence	r.m.s.	E_{vd}	From
K V I	0·15	7·9	IPAZ
M V N	0·71	0·0	2SNS
H V A	0·10	6·1	2PRK
T V D	0·63	−0·8	ICSE
L V L	0·20	−0·9	ICTF
. . .	. . .	. . .	. . .
L V A	0·07	−1·3	5CYT
L V E	0·07	−1·3	3TLN
F V S	0·07	−1·4	IECO
L V E	0·08	−1·6	3INS
L V T	0·03	−1·4	3HHB

Figure 2. An illustration showing how a segment of crambin is matched. (a) A central residue, i, is chosen at random. This defines a search segment of length L (here $L = 3$), which has a known sequence and a partially known conformation defined by any 3 known C^{α} co-ordinates. (b) Together these criteria define a template that is tested against a data base of refined protein structures to find data base segments that best match the template in terms of both conformation (measured by fit of inter-C^{α} distances) and sequence (measured by homology). (c) The 40 best matches are then checked using the additional criteria of atomic r.m.s. deviation (*RMS* in Å) and van der Waals' energy (*Evd* in kcal/mol). The list shown is taken from a real situation and has been sorted by decreasing pseudo-energy (0·1 *Evd* + *RMS*). The worst segments either have high energy values (for example, *Evd* = 7·9 and 6·1) or large deviations (for example, *RMS* = 0·71 Å and 0·63 Å). (d) One segment is selected at random from the N_c best segments on the list (here $N_c = 4$). All the segments with low pseudo-energy have low energy values and small r.m.s. deviations justifying such a random choice. (e) The atomic co-ordinates of the main or side-chain of the chosen segment are then added to the growing target structure. The process is repeated until all atoms have been positioned.

specification or prior modeling); and when building side-chains, all the main-chain atoms together with any known atoms of the central side-chain are used. For example, consider modeling a Tyr side-chain from a Phe. In this case the target structure will have input co-ordinates for all but the hydroxyl atom of Tyr. The modeled OH position must fit the other atoms of the side-chain and this can be ensured by giving these known side-chain atoms higher weights in the co-ordinate fitting procedure (by a factor of 4).

This fitting procedure provides the positional r.m.s. deviation between target and data base segment. In addition, the data base segment and the target structure are then in the same co-ordinate frame and atoms can be easily copied from the segment to the growing target. It is important to check that the copied atoms do not clash with atoms already in the target and this is done by calculating the van der Waals' energy between each data base segment and the target structure. As more segments are matched, more atoms become part of the target structure. Thus, the interaction of a data base segment and the growing target structure depends on the stage of building: segments can be added easily at the start and with increasing difficulty as the target structure grows to occupy more space.

In the energy calculation, atoms used from the target structure include all the located atoms (either from input or prior modeling) except those of residues $N_{start} - 1$ to $N_{end} + 1$, where the segment being modeled extends from N_{start} to N_{end}. Atoms used from the data base segment are selected as follows: when building the main chain, only the C^{α} atoms are included; and when building side-chains, the main-chain atoms of all residues and the side-chain atoms of the central residue are included. In both cases, interactions within the segment are ignored.

The van der Waals' energy, U_{vdw}, is calculated as a sum of 6–12 Lennard-Jones interactions between pairs of atoms closer than 6 Å. $U_{vdw} = \Sigma\varepsilon[(r_o/r)^{12} - 2(r_o/r)^6]$, where r, the interatomic distance, is set to 1 Å if the particular pair of atoms is closer than 1 Å. The parameters r_o and ε, taken from Levitt (1983*a*), are as follows: (r_o, ε) = (3·100,0·185) for O...O, (3·817,0·413) for N...N and (4·315,0·0738) for C...C and S...S; interactions of other atom pairs use the geometric means (r_o values are in Å and ε values are in kcal/mol; 1 cal = 4·184 J). Hydrogen atoms are excluded as these atoms are not generally available from X-ray crystallography. Provision for hydrogen bonding is made by setting the values (r_o,ε) = (2·9,3·0) for pair-wise interactions between C = O...N, C = O...OH, N...OH, OH...OH (OH denotes an oxygen atom in a hydroxyl group). This interaction has an energetic minimum of −3 kcal/mol at an atomic separation of 2·9 Å.

The co-ordinate deviation and van der Waals' energy of each data base segment are used to select a single segment that will provide the co-ordinates of the growing target structure. The most obvious scheme is to select the data base segment with both the lowest values of the r.m.s. deviation (*RMS*) and van der Waals' energy (*Evd*). To do this, the r.m.s. deviation and energy are put onto a single scale in a pseudo-energy term: $U = 0{\cdot}1\ Evd + RMS$. This makes a 1 kcal/mol change in energy equivalent to a 0·1 Å change in r.m.s. deviation.

The example in Fig. 2 shows the *RMS* and *Evd* values for some of the 40 selected segments in the modeling of one residue of crambin (CRN). The different segments come from different data base proteins as indicated by the protein identifier (the **From** column). It is interesting that in this example the best data base segments share a characteristic sequence pattern, with the 1st position favoring L and the 3rd position being E, A, S or T.

(iv) *Selecting a best match*

Rather than choosing the data base segment with the best value of the pseudo-energy, the choice is made less determinate. The segment, i, with pseudo-energy, U_i, is chosen with a probability proportional to exp $(-\beta U_i)$, where the choice is limited to those N_c segments with the lowest values of U_i. When $\beta = 0$, the selected segment is chosen randomly from the best N_c segments. The β and N_c parameters do influence the results and ranges of values have been explored (see Table 10).

(v) *Copy co-ordinates from segment to target*

Having selected a data base segment, all that remains is to copy certain atomic co-ordinates to the target structure. When building the main-chain, all the unknown main-chain atoms of the central residue, i, are copied to the target. When building the side-chain, all the unknown side-chain atoms of the central residue are copied to the target. Once an atom is copied to the target structure it is treated as being "known" and can be used for subsequent modeling. Other schemes that involved copying more atoms to the target, for example, the main-chain atoms of the central residue and the atoms of the preceding residue, worked less well. In certain circumstances, it may be justified to use data base co-ordinates even when the co-ordinates of the particular atom are known: for example, if the initial co-ordinates are known to be subject to large errors.

(c) *Randomization and averaging*

Several aspects of the modeling scheme described above are controlled by random numbers: the central residue of both the main and side-chain segments is chosen at random and the particular data base segment is selected with a Boltzmann probability. Randomization is done to avoid having to make choices that could not be justified. For example, where is the best place to start the modeling? Is the fragment with the lowest pseudo-energy always the best? By building segments in random order and making random choices of the best segment at each stage, systematic errors related to the incremental nature of the process are avoided. The price paid for this is that the modeling is not unique and many different models can be built by the same procedure. How is the best model to be chosen? Several schemes suggest themselves: for example, one could take the model with the lowest energy, or the most buried hydrophobic groups.

Here a much simpler scheme is adopted. A number of independent models are generated and used to calculate mean co-ordinates. Initially, this was done to determine the reproducibility of the modeling procedure: those parts of the structure that are more similar in the different models will have a smaller co-ordinate variance and may be expected to be more like the X-ray structure. The mean co-ordinates are needed to calculate the variance of each atom. Surprisingly, the r.m.s. deviation of these mean co-ordinates from the known X-ray structure is always found to be significantly smaller than the r.m.s. deviation of any of the individual models. This led to the use of averaging as a standard part of the procedure.

Averaging is not quite straightforward, in that certain side-chains are intrinsically degenerate in their atom labeling. For example, the pairs of atoms OD1 and OD2 of Asp, CD1 and CD2, and CE1 and CE2 of Phe are equivalent. Here equivalent groups of atoms are exchanged to make each model deviate least from the mean model calculated for the structures modeled so far. This exchange can be thought of as a rotation of 180° about χ_2 of Asp, Phe and Tyr, and about χ_3 of Glu. This procedure is always done without using any knowledge of the structure of the target protein as this would bias the r.m.s. deviation towards the X-ray structure. Incorrectly labeled methyl groups in Val and Leu side-chains are also checked by exchanging CG1 and CG2 of Val and CD1 and CD2 of Leu if necessary.

(d) *Energy refinement*

One drawback of co-ordinate averaging is that a model calculated as the mean of several sets of co-ordinates can have poor stereochemistry. Correct stereochemistry is enforced by minimization of the energy using the program ENCAD (Levitt, 1983*a*). The program makes use of a standard interatomic potential energy function, consisting of bond stretching, angle bending, torsional, non-bonded and electrostatic interactions. Polar hydrogen atoms are included to allow accurate electrostatic hydrogen bonds. Such minimization also corrects the poor stereochemistry that results from closing the gaps between disjoint segments.

Because energy minimization is intended to enforce good stereochemistry without moving atoms unduly, the C^α atoms are sometimes restrained to their initial positions. The restraint term, which is added to the potential energy minimized by ENCAD, is of the unusual form $U = K\ (drms/\delta)^8$, where the force-constant K is 1 kcal/mol Å, and δ, which determines the maximum tolerated value of *drms*, is 0·05 Å for "tight" minimization and 0·3 Å for "loose" minimization. The value of *drms* is given by $drms^2 = \Sigma(r_{ij}{}^2\text{-}R_{ij}{}^2)^2/4\ \Sigma R_{ij}{}^2$, where r_{ij} is the separation of atoms i and j in the current structure, R_{ij} is the corresponding distance in the reference X-ray structure, and the summation includes all atom pairs with $R_{ij} > 10$ Å. Expanding $(r_{ij}{}^2\text{-}R_{ij}{}^2)^2$ as $(r_{ij}+R_{ij})^2\ (r_{ij}\text{-}R_{ij})^2$ and approximating $(r_{ij}+R_{ij})$ as $2R_{ij}$ gives $drms^2 = \Sigma R_{ij}{}^2(r_{ij}\text{-}R_{ij})^2/\Sigma R_{ij}{}^2$. This shows that *drms* is the r.m.s. deviation of the long interatomic distances, weighted by the square of the distance in the X-ray structure (long distances have much more weight than shorter distances). This form is chosen as it does not require use of square-roots, which are computationally inefficient. The use of the 8th power in the calculation of the restraint energy means that when *drms* exceeds δ, the restraint energy rises rapidly, whereas for smaller values the restraint energy is negligible. The use of only the longer distances in the calculation means that local geometry is unaffected while the overall shape of the chain is strongly restrained. When modeling from a subset of the C^{α} atoms, the restraint is applied to only those C^{α} atoms that are assumed to be known from the X-ray structure. Minimization is done for a total of 600 energy function evaluations using the quadratically convergent conjugate gradient method. When no restraints are used, the minimization is referred to as free minimization.

The same energy minimization program also provides in-depth analysis of the refined atomic co-ordinates by tabulating r.m.s. deviations of atom classes, residues and chains, hydrogen bonds, accessible surface areas and torsion angles. For the analysis presented here, most use is made of r.m.s. deviations of atomic positions and torsion angles.

(e) *Computing requirements*

The computer requirements of segment match modeling are modest. Building all 10 independent models takes 6 s/residue and does not depend on the number of atoms used as guide points (unless atoms are provided for entire residues that then do not need to be re-built). Restrained energy minimization lasting 600 steps takes 8 s/residue. Thus, the total time required to model a protein is 14 s/residue (a 200-residue protein requires 45 min on the Silicon Graphics 4D/25 Personal Iris workstation).

3. Results

Segment match modeling has been tested on eight well-refined proteins. For test purposes, each of these proteins is taken from the data base (see Table 1) but all data for the particular protein are omitted when it is being modeled. The proteins selected are as follows (the Protein Data Base name and reference is given in parentheses): CRN, crambin (1CRN, Hendrickson & Teeter, 1981); PTI, trypsin inhibitor (5PTI, Wlodawer *et al.*, 1984); CTF, C-terminal fragment of L7/L12 ribosomal protein (1CTF, Leijonmark & Liljas, 1987); RNS, ribonuclease A (1RN3, Borkakoti *et al.*, 1982); LYZ, lysozyme (1LZ1, Artymiuk & Blake, 1981); FXN, flavodoxin (4FXN, Smith *et al.*, 1977); TLN, thermolysin (3TLN, Holmes & Matthews, 1982); and APP, penicillopepsin acid protease (1APP, James & Sielecki, 1983). These structures contain between 46 and 323 amino acid residues.

(a) *Overall accuracy of segment match modeling*

The results given in Table 2 show that the models constructed from the C^{α} co-ordinates are surprisingly accurate. In the structures produced by "tight" energy minimization, the overall r.m.s. deviation is 0·42 Å for main-chain atoms, 1·78 Å for side-chain atoms and 1·26 Å for all non-hydrogen atoms. Different proteins are modeled more or less well: the results are consistently best for CTF and worst for PTI. Large proteins are modeled as well as small proteins: CTF, a small protein has the lowest all-atom r.m.s. deviation of 0·93 Å, whereas APP, the largest protein, has the second lowest all-atom r.m.s. of 1·00 Å. The main-chain r.m.s. deviation shows less variation between protein structures than does the side-chain or all-atom r.m.s. deviation.

Table 2 shows the r.m.s. deviations of models produced by segment match modeling without employing energy minimization. Ten independent models are built for each protein using different random numbers to control the order of building and the choice of best-matching segments. The different models vary considerably: the overall all-atom r.m.s. deviation of the worst of the ten models (1·81 Å) is 20% higher than that of the best of the ten models (1·49 Å). More surprising, the r.m.s. deviation of the mean model (1·19 Å), calculated as the mean of the atomic co-ordinates or the ten models, is much better than the best of the ten models. This comparison is not strictly fair, since each model had reasonably good stereochemistry, whereas the mean model can have very poor stereochemistry. The models generated by energy minimization from these mean co-ordinates with tight restraints did have excellent stereochemistry and an all-atom r.m.s. deviation (1·26 Å) that is only a little higher than that of the mean model. Thus, the procedure of generating several independent models, calculating a mean model and then refining the stereochemistry provides an unexpectedly powerful way of improving the accuracy of the models. It is important to note that while the mean of the ten models can always be calculated and refined, the best of the ten models can only be found by knowing the X-ray structure. The difference in the r.m.s. deviation may seem small (for example, for side-chains it is 2·40 Å for an average model and 1·78 Å for the refined mean model), but these differences are significant. The lowering of the r.m.s. deviation caused by averaging implies that the positions of some of the atoms in any one of the ten models are inaccurate, whereas those positions that occur in several of the models are more likely to be correct.

(b) *Root-mean-square deviations of models*

Table 3 presents a more detailed examination of the r.m.s. deviations of the structures generated by "tight" minimization of the mean models. Each of these structures has good internal geometry in that bond lengths, bond angles and torsion angles are close to their standard values and there are no short interatomic contacts other than hydrogen bonds. Main and side-chain r.m.s. deviations are given for different classes of amino acids and for each individual protein. These measures of deviation are also given for three sub-classes of atoms: (1) those atoms that are the most buried from solvent (Lee & Richards, 1971); (2) those atoms that form most main-chain hydrogen bonds; and (3) those atoms that are most reproducibly modeled (have the lowest variance or standard deviation for the 10 independent models). In each case, the threshold for inclusion into the class is chosen to give approxi-

Table 1

The 76 high-resolution structures used as the data base of known proteins

Res[a] (Å)	Name	Date	Full name	Reference[b]
1·00	5PTI	10/84	Trypsin inhibitor	Wlodawer *et al.* (1984)
1·20	1CSE	6/88	Subtilisin Carlsberg/Eglin	Bode *et al.* (1986)
1·34	1UTG	3/89	Uteroglobin	Morize *et al.* (1987)
1·34	4PTP	4/88	Trypsin (DIP inhibited)	Chambers & Stroud (1977)
1·37	1PPT	1/81	Avian pancreatic polypeptide	Blundell *et al.* (1981)
1·38	1NXB	8/80	Neurotoxin B	Tsernoglou & Petsko (1977)
1·40	1ECO	3/79	Erythrocruorin (CO)	Steigemann & Weber (1979)
1·40	1MBD	8/81	Myoglobin (deoxy)	Phillips (1980)
1·40	1RDG	10/84	Rubredoxin	Frey *et al.* (1987)
1·40	3EBX	1/88	Erabutoxin B	Smith *et al.* (1988)
1·45	1RN3	10/81	Ribonuclease A	Borkakoti *et al.* (1982)
1·50	1CRN	5/81	Crambin	Hendrickson & Teeter (1981)
1·50	1LZ1	10/84	Lysozyme	Artymiuk & Blake (1981)
1·50	2OVO	6/85	Ovomucoid 3rd domain	Bode *et al.* (1985)
1·50	2PRK	11/87	Proteinase K	Betzel *et al.* (1988)
1·50	2SGA	1/83	Proteinase A	Sielecki *et al.* (1979)
1·50	2SNS	5/82	Staphylococcal nuclease	Cotton *et al.* (1979)
1·50	3INS	10/88	Insulin	Wlodawer *et al* .(1989)
1·50	5CYT	1/88	Cytochrome *c*	Takano (1984)
1·54	3GRS	2/88	Glutathione reductase	Karplus & Schulz (1987)
1·54	5CPA	5/82	Carboxypeptidase A	Rees *et al.* (1983)
1·55	1PAZ	6/88	Pseudoazurin	Petratos *et al.* (1987)
1·60	1GCR	8/85	γ-II Crystallin	Wistow *et al.* (1983)
1·60	1PCY	8/80	Plastocyanin (Cu2+)	Guss & Freeman (1983)
1·60	2RHE	6/83	Bence-Jones dimer	Furey *et al.* (1983)
1·60	3TLN	2/82	Thermolysin (native)	Holmes & Matthews (1982)
1·60	451C	7/81	Cytochrome *c*551 (reduced)	Matsuura *et al.* (1982)
1·63	2CPP	4/87	Cytochrome P450CAM	Poulos *et al.* (1987)
1·65	2WRP	12/87	Trp repressor (orthorombic)	Zhang *et al.* (1987)
1·65	3EST	5/76	Elastase	Meyer *et al.* (1988)
1·65	9PAP	3/86	Papain (oxidized Cys25)	Kamphius *et al.* (1984)
1·67	2CCY	8/85	Cytochrome *c'*	Finzel *et al.* (1985)
1·67	5CHA	1/85	α-Chymotrypsin	Blevins & Tulinsky (1985)
1·70	1ALC	8/89	α-Lactalbumin	Acharya *et al.* (1989)
1·70	1BP2	4/81	Phospholipase A2	Dijkstra *et al.* (1981)
1·70	1CTF	9/86	L7/L12 50 S ribosomal protein	Leijonmarke & Liljas (1987)
1·70	2ACT	11/79	Actinidine (sulfhydryl proteinase)	Baker & Dodson (1980)
1·70	2ALP	3/85	α-Lytic protease	Fujinaga *et al.* (1985)
1·70	2LZM	8/86	Lysozyme	Weaver & Matthews (1987)
1·70	2MHR	4/87	Myohemerythrin	Sherrif *et al.* (1987)
1·70	8DFR	5/89	Dihydrofolate reductase	Matthews *et al.* (1985)
1·74	3HHB	3/84	Hemoglobin (deoxy)	Fermi *et al.* (1984)
1·80	1GOX	6/89	Glycolate oxidase	Lindquist (1989)
1·80	1SN3	12/82	Scorpion neurotoxin (variant 3)	Almassy *et al.* (1983)
1·80	1TON	6/87	Tonin	Fujinaga & James (1987)
1·80	1UBQ	1/87	Ubiquitin	Vijay-Kumar *et al.* (1987)
1·80	2APP	1/83	Acid proteinase	James & Sielecki (1983)
1·80	2AZA	10/86	Azurin	Baker (1988)
1·80	2CDV	6/85	Cytochrome *c*3	Higuchi *et al.* (1984)
1·80	2PAB	9/77	Prealbumin	Blake *et al.* (1978)
1·80	3SGB	1/83	Proteinase B (OMTKY3)	Read *et al.* (1983)
1·80	3WGA	3/86	Agglutinin (iso)	Wright (1987)
1·80	4FXN	12/77	Flavodoxin (red)	Smith *et al.* (1977)
1·84	1HNE	4/89	Elastase	Navia *et al.* (1989)
1·85	1CPV	8/74	Ca^{2+}-binding parvalbumin	Moews & Kretsinger (1975)
1·90	2CA2	2/89	Carbonic anhydrase II / SCN	Eriksson *et al.* (1988)
1·90	2FB4	4/89	IGG1 Fab (lambda) KOL	Marquart *et al.* (1980)
1·90	3RP2	9/84	Proteinase II	Remington *et al.* (1988)
1·90	4FD1	6/88	Ferredoxin	Stout (1989)
2·00	1ACX	12/82	Actinoxanthide	Pletnev *et al.* (1982)
2·00	1GP1	6/85	Glutathione peroxidase	Epp *et al.* (1983)
2·00	1HIP	4/75	High potential iron protein	Carter *et al.* (1974)
2·00	1HMQ	2/83	Hemerythrin (met)	Stenkamp *et al.* (1983)
2·00	1HOE	1/89	α-Amylase inhibitor HOE467A	Pflugrath *et al.* (1986)
2·00	1I1B	12/89	Interleukin 1B	Finzel *et al.* (1989*a*)
2·00	1MLT	8/81	Melittin	Terwilliger & Eisenberg (1982)
2·00	1R69	12/88	Repressor (1-69)	Mondragon *et al.* (1989)
2·00	2BC5	12/77	Cytochrome *b*5 (oxidized)	Mathews *et al.* (1972)
2·00	2CI2	9/88	Chymotrypsin inhibitor 2	McPhalen & James (1987)
2·00	2CNA	4/75	Concanavalin A	Reeke *et al.* (1975)

Table 1 *(continued)*

Res[a] (Å)	Name	Date	Full name	Reference[b]
2·00	2LH1	4/82	Leghemoglobin (acetate met)	Arutyunyan *et al.* (1980)
2·00	2RSP	10/89	Rous sarcoma virus protease	Miller *et al.* (1989)
2·00	2SOD	3/80	Superoxide dismutase	Tainer *et al.* (1982)
2·00	3FAB	9/81	Fab' (lambda) NEW	Saul *et al.* (1978)
2·00	4TNC	9/87	Troponin C	Satyshur *et al.* (1988)
2·00	6LDH	11/87	Lactate dehydrogenase	Abad-Zapatero *et al.* (1987)

[a] Res is the resolution of the diffraction data used in the structure determination; for convenience, the structures are sorted by decreasing resolution but the modeling procedure is independent of this order.
[b] References are taken from the Brookhaven Protein Data Bank co-ordinate files.

mately equal numbers of atoms in and out of the class for each amino acid.

For the main-chain, the atoms that are most involved in hydrogen bonds (the HB class) or are most reproducible (the Rep class) have smaller overall r.m.s. deviations (0·30 Å as opposed to 0·40 Å). For the side-chains, the atoms that are most reproducibly modeled have significantly smaller overall r.m.s. deviations (1·47 Å as opposed to 1·78 Å). The main-chain r.m.s. deviation is worst for Gly (0·56 Å) and Trp (0·61 Å) residues. This finding for Gly is understandable as this amino acid has greater conformational freedom than other amino acids. The side-chain r.m.s. deviation is the

Table 2
Summary of r.m.s. deviations for modeled proteins

		r.m.s. deviation (Å)				Minimized		
Protein	N_{res}[a]	Worst[b]	Best	Ave	Mean	Tight	Loose	Free
A. *Main-chain atoms*								
CRN	46	0·69	0·61	0·64	0·61	0·56	0·74	0·92
PTI	58	0·64	0·45	0·53	0·46	0·51	0·64	0·81
CTF	68	0·49	0·35	0·43	0·30	0·29	0·42	0·59
RNS	124	0·61	0·48	0·52	0·40	0·41	0·53	0·86
LYZ	129	0·63	0·49	0·58	0·45	0·39	0·54	0·76
FXN	138	0·60	0·49	0·54	0·50	0·44	0·54	0·78
TLN	316	0·55	0·50	0·53	0·46	0·38	0·64	0·88
APP	323	0·51	0·44	0·47	0·40	0·37	0·55	0·76
Overall[c]		0·59	0·48	0·53	0·45	0·42	0·58	0·80
B. *Side-chain atoms*								
CRN	46	2·49	2·09	2·17	1·33	1·57	1·60	1·71
PTI	58	2·95	2·16	2·70	1·97	2·43	2·42	2·60
CTF	68	2·19	1·79	1·97	1·31	1·37	1·43	1·52
RNS	124	2·53	2·21	2·41	1·82	2·00	2·06	2·26
LYZ	129	2·82	2·30	2·53	1·66	1·59	1·74	1·86
FXN	138	2·57	2·22	2·39	1·83	1·91	1·95	2·07
TLN	316	2·56	2·33	2·43	1·83	1·94	2·03	2·17
APP	323	2·33	1·99	2·20	1·48	1·42	1·53	1·65
Overall		2·56	2·14	2·40	1·65	1·78	1·85	1·98
C. *All atoms*								
CRN	46	1·69	1·32	1·51	0·99	1·12	1·19	1·33
PTI	58	2·11	1·58	1·92	1·41	1·73	1·75	1·91
CTF	68	1·59	1·23	1·35	0·90	0·93	1·00	1·10
RNS	124	1·79	1·57	1·79	1·29	1·41	1·47	1·68
LYZ	129	2·03	1·65	1·83	1·21	1·15	1·28	1·42
FXN	138	1·83	1·59	1·71	1·33	1·37	1·42	1·55
TLN	316	1·82	1·66	1·73	1·31	1·37	1·48	1·63
APP	323	1·61	1·38	1·53	1·04	1·00	1·11	1·25
Overall		1·81	1·49	1·67	1·19	1·26	1·34	1·48

[a] N_{res} is the number of residues in the protein and shows that we use proteins of all sizes.
[b] Each protein is modeled 10 times using different random numbers. Worst, Best and Ave refer to the highest, lowest and average r.m.s. deviation from the X-ray structure for this set of 10 models. Mean refers to the model formed by taking the mean of each Cartesian co-ordinate for the 10 models. As this Mean structure can have very poor stereochemistry, it is subjected to energy minimization done in 3 ways: Tight, with a strong restraint holding the C^{α} atoms to the X-ray C^{α} positions (δ = 0·05 Å); Loose, with a weaker restraint to these same positions (δ = 0·30 Å); and Free, no restraint whatsoever. In each case, 600 steps of minimization were done.
[c] The Overall value is the average of the corresponding values in the 8 proteins.

Table 3

Root-mean-square deviations of classes of amino acids

Amino acid	Number of cases	r.m.s. deviation (Å)[a] Main-chain Any	Bur	HB	Rep	Side-chain Any	Bur	HB	Rep
A. *Special*									
Gly	120	0·56	0·53	0·45	0·54	—[b]	—	—	—
Pro	39	0·36	0·37	0·42	0·28	0·35	0·35	0·35	0·34
Cys	33	0·27	0·30	0·27	0·19	1·06	1·16	0·94	0·84
B. *Non-polar*									
Ala	110	0·32	0·21	0·17	0·17	0·39	0·18	0·16	0·17
Val	84	0·39	0·18	0·17	0·19	0·70	0·51	0·45	0·59
Ile	63	0·30	0·17	0·16	0·18	1·28	1·12	1·07	0·81
Leu	65	0·28	0·32	0·33	0·21	0·85	0·71	0·85	0·70
Met	14	0·18	0·15	0·13	0·15	1·15	1·00	0·83	1·08
C. *Aromatic*									
His	16	0·51	0·65	0·32	0·27	1·72	1·46	2·22	1·48
Phe	46	0·22	0·24	0·24	0·21	2·12	1·36	1·56	1·31
Tyr	63	0·24	0·24	0·19	0·22	1·80	1·15	1·06	1·39
Trp	14	0·61	0·74	0·16	0·27	3·27	2·93	1·55	1·80
D. *Polar*									
Ser	106	0·27	0·29	0·28	0·25	1·15	1·16	1·26	1·03
Thr	85	0·52	0·44	0·44	0·41	1·15	1·09	1·06	0·87
Asn	69	0·44	0·54	0·25	0·32	1·83	1·80	1·64	1·64
Gln	54	0·41	0·22	0·24	0·29	1·96	1·83	1·89	1·72
E. *Charged*									
Asp	77	0·45	0·21	0·38	0·38	1·38	1·22	1·27	1·16
Glu	51	0·41	0·31	0·41	0·42	1·88	1·84	1·86	1·72
Lys	55	0·52	0·57	0·17	0·38	2·49	2·22	2·12	2·31
Arg	39	0·28	0·23	0·22	0·23	2·78	2·81	2·86	2·72
F. *Overall*[c]									
All	1203	0·40	0·36	0·30	0·31	1·78	1·57	1·52	1·47
No KR	1109	0·40	0·35	0·31	0·31	1·62	1·37	1·31	1·25
No FYWKR	986	0·41	0·35	0·32	0·32	1·37	1·29	1·31	1·18
G. *Individual proteins*									
CRN	46	0·56	0·94	0·40	0·22	1·57	2·25	0·99	0·69
PTI	58	0·51	0·47	0·40	0·54	2·43	1·21	1·11	1·57
CTF	68	0·29	0·39	0·31	0·31	1·37	1·56	1·48	1·38
RNS	124	0·41	0·30	0·25	0·26	2·00	1·47	1·76	1·84
LYZ	130	0·39	0·36	0·35	0·29	1·59	1·37	1·36	1·47
FXN	138	0·44	0·38	0·37	0·34	1·91	1·71	1·67	1·63
TLN	316	0·38	0·34	0·27	0·31	1·94	1·86	1·63	1·51
APP	323	0·37	0·30	0·25	0·29	1·42	1·25	1·35	1·20

Different types of amino acid are grouped together. Results are given for the structures after tight minimization.

[a] The main-chain and side-chain r.m.s. deviations for each amino acid from the respective X-ray structures are calculated for 4 classes of atoms as follows: Any includes any atom; Bur includes the residues that are most buried (have solvent accessibility less than 70 Å^2, Lee & Richards, 1971); HB includes the residues that are most involved in main-chain hydrogen bonds (at least 1 such hydrogen bond/residue); Rep includes the atoms of the residues that are most reproducibly modeled (have smallest standard deviations between the different independent models). These thresholds are chosen so that about half the atoms are in the class for each type of amino acid.

[b] There are no residues in this class.

[c] Overall is averaged over all the residues in these proteins. This value is slightly different from the overall value given in Table 2 as that value is averaged over the 8 proteins. No KR indicates all residues except Lys and Arg, whereas No FYWKR indicates all residues except Phe, Tyr, Trp, Lys and Arg.

worst for Trp and Arg residues. This is partially explained by the fact that both side-chains have many atoms. Note, however, that the three aromatic side-chains, His, Phe and Tyr, all have at least as many atoms as Arg yet have much lower side-chain r.m.s. deviations. Arg may be badly modeled since its conformation will be affected by solvent and crystal contacts. The side-chain r.m.s. deviations are below 1 Å for Pro, Ala, Val and Leu residues. The main-chain atoms, N, CA, C and O, have overall r.m.s. deviations of 0·22, 0·12, 0·30 and 0·73 Å, respectively, showing that modeling of the oxygen atoms is much less accurate than for other main-chain atoms.

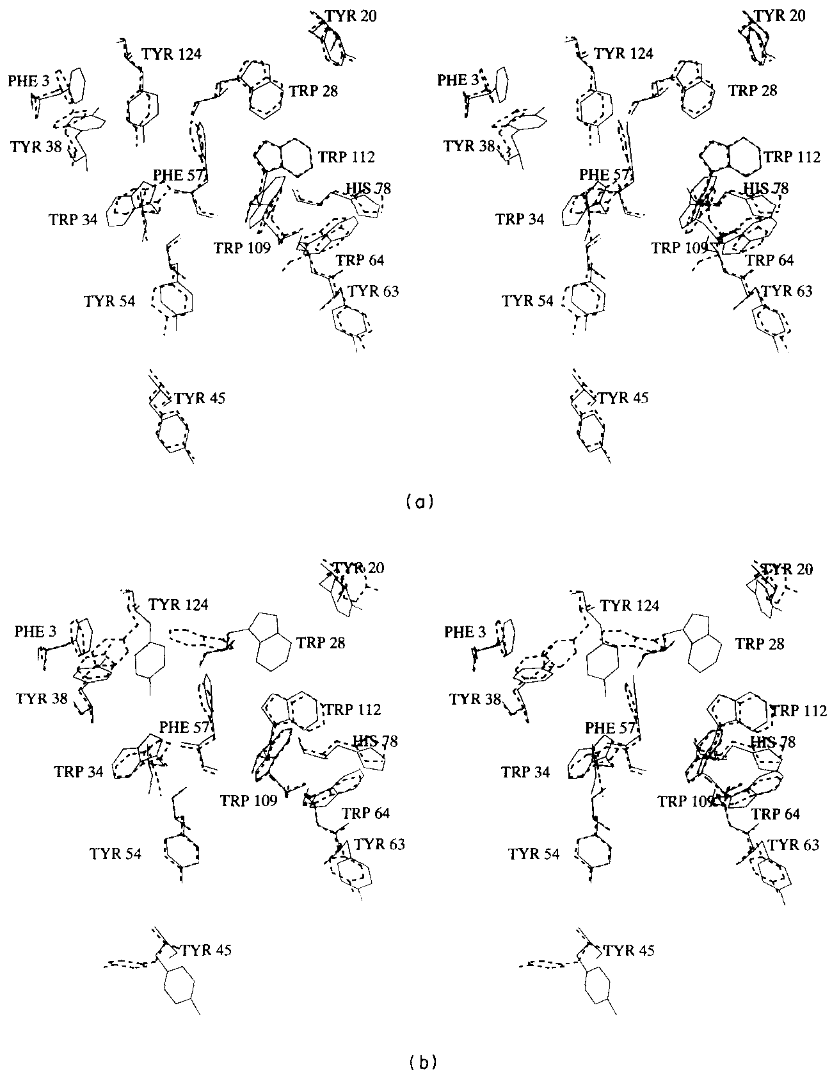


Figure 3. (a) Stereo view comparing the conformation of the 14 aromatic residues (His, Phe, Trp and Tyr) in the lysozyme (LYZ) X-ray structure (continuous line) and the corresponding structure built by segment match modeling using all the C^{α} positions (broken line). It is clear that almost all of these side-chains are correctly predicted. Closer inspection shows that in 2 cases Trp residues occupy similar volumes but are flipped 180° about the C^{β}–C^{γ} bond (Trp64 and Trp109). The r.m.s. deviation of all side-chain atoms is 1·59 Å. (b) Stereo view comparing the conformation of all aromatic residues in the X-ray structure (continuous line) and the structure built by segment match modeling using every 2nd C^{α} position (broken line). The agreement is less good with errors in the positions of Trp28, Tyr124 and Tyr45. The r.m.s. deviation of all side-chain atoms is 2·20 Å. Note that Trp64 and Trp109 are now modeled correctly. This and the other molecular drawings were made on an Apple Macintosh computer using the programs MacImdad (Molecular Applications Group) and Canvas (Deneba Software).

The r.m.s. deviations, even when sub-divided into classes of atoms, are of limited value and must be supplemented by other measures. The most objective scheme is to show the co-ordinates themselves in stereoscopic drawings comparing the X-ray and modeled structures. Here, attention is focused on the aromatic residues, as (1) showing all residues obscures important detail, (2) the aromatic side-chains are large, (3) errors in the χ_1 or χ_2 torsion angles are immediately apparent in the drawing and (4) this class is not modeled particularly well (see Table 3). Figure 3(a) shows these residues for the 130-residue protein lysozyme (LYZ). It is immediately obvious that segment match modeling works extremely well for this protein and that the aromatic side-chains are well-predicted. Figure 4 shows similar comparisons for another protein, flavodoxin (FXN), which has more errors. The r.m.s. of all side-chain atoms is 1·59 Å and 1·91 Å for LYZ and FXN, respectively (Table 3).

One of the most useful measures of modeling would be a simple binary decision: is the particular residue right or wrong? What is really meant by this question is whether the residue is close enough to its X-ray position to be useful for other studies. Here, a wrong side-chain is defined as one whose r.m.s. deviation exceeds 2 Å and Table 4 gives the number of wrongly modeled side-chains in the eight test proteins. Comparison of the results for the aromatic residues with structures drawn in Figures 3 and 4

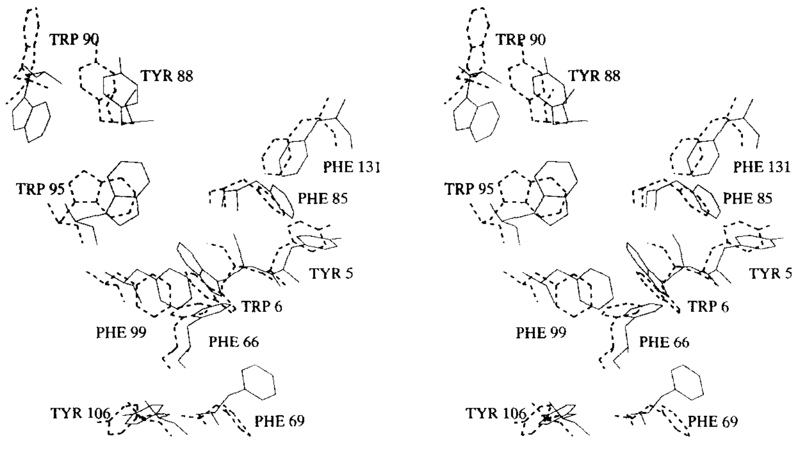


Figure 4. Stereo view comparing the X-ray (continuous line) and modeled (broken line) positions of all aromatic residues (His, Phe, Trp and Tyr) in the protein flavodoxin (FXN). The obvious discrepancies involve Phe69 and Trp90. Closer inspection shows that Trp6 is shifted and Trp95 is flipped about χ_2 so that both these residues are also considered as errors (see Table 4). The r.m.s. deviation of all side-chain atoms is 1·91 Å.

indicates that the 2 Å threshold corresponds well to a subjective impression of goodness of fit (for LYZ in Fig. 3(a), the error rate is 2/14; for FXN in Fig. 4, it is 4/11). The error rate depends on the amino acids involved. The small and non-polar side-chains are almost all correct, with the exception of Ile, which has an error rate of 22%. The overall error rate for the aromatic residues is 25%, with PTI having more errors than expected and LYZ fewer errors. The polar side-chains are modeled less accurately than non-polar side-chains of comparable size. It is not clear why this should occur since the van der Waals' potential includes an approximate hydrogen bonding term. Interactions between these side-chains and the solvent may have a major influence on their conformations. The error rate for

Table 4

Incorrect side-chains for different amino acids

		Fraction incorrect[a]								
Amino acid	No. atoms	CRN	PTI	CTF	RNS	LYZ	FXN	TLN	APP	Over-all
A. *Special*		0/15	0/16	0/7	0/15	0/21	1/20	0/44	0/54	1/192
Gly	0	0/4	0/6	0/6	0/3	0/11	0/14	0/36	0/40	0/120
Pro	4	0/5	0/4	0/1	0/4	0/2	0/3	0/8	0/12	0/39
Cys	2	0/6	0/6	0/0	0/8	0/8	1/3	0/0	0/2	1/33
B. *Non-polar*		1/13	0/12	1/32	2/30	0/38	4/44	6/86	5/81	19/336
Ala	1	0/5	0/6	0/15	0/12	0/14	0/6	0/28	1/24	1/110
Val	3	0/2	0/1	0/8	1/9	0/9	1/10	0/22	0/23	2/84
Ile	4	1/5	0/2	0/2	1/3	0/5	3/15	5/18	4/13	14/63
Leu	4	0/1	0/2	1/7	0/2	0/8	0/8	1/16	0/21	2/65
Met	4	0/0	0/1	0/0	0/4	0/2	0/5	0/2	0/0	0/14
C. *Aromatic*		0/3	4/8	0/1	3/13	2/14	4/11	12/59	10/40	35/139
His	6	0/0	0/0	0/0	0/4	0/1	0/0	2/8	1/3	3/16
Phe	7	0/1	3/4	0/1	0/3	0/2	1/5	2/10	5/20	11/46
Tyr	8	0/2	1/4	0/0	3/6	0/6	0/3	6/28	3/14	13/63
Trp	10	0/0	0/0	0/0	0/0	2/5	3/3	2/3	1/3	8/14
D. *Polar*		2/11	0/8	1/4	5/42	7/27	2/23	12/83	17/116	46/314
Ser	2	0/2	0/1	0/2	0/15	0/6	0/8	0/26	3/46	3/106
Thr	3	1/6	0/3	0/1	1/10	0/5	0/5	3/25	2/30	7/85
Asn	4	1/3	0/3	1/1	3/10	2/10	2/8	4/19	5/15	18/69
Gln	5	0/0	0/1	0/0	1/7	5/6	0/2	5/13	7/25	18/54
E. *Charged*		1/4	9/14	6/24	13/24	10/30	18/40	21/54	5/32	81/222
Asp	4	0/1	0/2	0/4	2/5	1/8	3/9	4/25	4/23	14/77
Glu	5	0/1	2/2	1/9	1/5	1/3	9/19	3/8	0/4	17/51
Lys	5	0/0	2/4	4/10	7/10	1/5	5/10	8/11	1/5	28/55
Arg	6	1/2	3/6	1/1	3/4	7/14	1/2	6/10	0/0	22/39
F. *All*		4/46	11/58	8/68	23/124	19/130	29/138	51/316	37/323	182/1203

[a] A side-chain is considered incorrect if its r.m.s. deviation exceeds 2·0 Å. Results are given for the structures after tight minimization.

Table 5

Fraction of incorrect side-chains for different classes of amino acids

	Fraction incorrect					
	Buried		H bonds		Reproducible	
Amino acid	Yes	No	Yes	No	Yes	No
A. *Special*	1/97	0/95	1/94	0/98	0/114	1/78
Gly	0/60	0/60	0/59	0/61	0/77	0/43
Pro	0/20	0/19	0/19	0/20	0/20	0/19
Cys	1/17	0/16	1/16	0/17	0/17	1/16
B *Non-polar*	6/170	13/166	5/164	14/172	3/174	16/162
Ala	0/56	1/54	0/54	1/56	0/57	1/53
Val	0/42	2/42	0/41	2/43	0/44	2/40
Ile	6/32	8/31	4/31	10/32	2/32	12/31
Leu	0/33	2/32	1/32	1/33	1/33	1/32
Met	0/7	0/7	0/6	0/8	0/8	0/6
C. *Aromatic*	12/71	23/68	11/66	24/73	10/74	25/65
His	1/8	2/8	2/7	1/9	1/9	2/7
Phe	2/23	9/23	3/22	8/24	2/24	9/22
Tyr	5/33	8/30	4/31	9/32	3/32	10/31
Trp	4/7	4/7	2/6	6/8	4/9	4/5
D. *Polar*	22/161	24/153	19/154	27/160	19/163	27/151
Ser	1/53	2/53	1/52	2/54	1/54	2/52
Thr	5/43	2/42	3/42	4/43	3/44	4/41
Asn	7/35	11/34	8/34	10/35	8/36	10/33
Gln	9/30	9/24	7/26	11/28	7/29	11/25
E. *Charged*	32/113	49/109	35/108	46/114	36/114	45/108
Asp	5/39	9/38	5/38	9/39	5/40	9/37
Glu	7/26	10/25	9/25	8/26	7/26	10/25
Lys	9/28	19/27	12/27	16/28	12/28	16/27
Arg	11/20	11/19	9/18	13/21	12/20	10/19
F. *All*	73/612	109/591	71/586	111/617	68/639	114/564

the charged side-chains is very high, especially for Arg, whose conformation is wrong more often than right. Overall, there are 182 wrong side-chains out of 1203 for an error rate of 15%. If the charged side-chains are omitted, the error rate drops to 10% (101 wrong out of 981).

It is instructive to compare the number of incorrect side-chains in the sub-class of residues that is most buried with the sub-class that is least buried (the remaining residues). Similar comparisons can be done using the extent of hydrogen bonding and the modeling reproducibility as criteria to define sub-classes of residues. Results for the non-polar amino acids are most dramatic (see Table 5). Of the 174 side-chains modeled with the smallest variance, only three are judged incorrect. For the other 162 non-polar side-chains, there are 16 errors. Thus, the error rate (calculated as (16/162)/(3/174)) is almost six times lower for those side-chains modeled most reproducibly relative to those side chains modeled least reproducibly. The criteria of being most buried or most hydrogen bonded also gave lower error rates (by 2·2-fold and 2·7-fold, respectively). This trend is seen also for the aromatic amino acids where the error rates are 2·0, 2·0 and 2·8 times lower for buried, hydrogen bonded and reproducibly modeled side-chains, respectively. For non-polar and aromatic amino acids taken together, there are a total of 13 incorrect side-chains out of the 248 modeled most reproducibly, giving an error rate of 5%. For these amino acids, modeling reproducibility is a better criterion than being buried or hydrogen bonded. It is noteworthy that reproducibility of every atom is obtained directly from segment match modeling and without any reference to the known structure, whereas burial and hydrogen bonding are derived from the X-ray co-ordinates. Reproducible side-chains are simply those that are modeled in the same way in the independent runs of the program. For the polar and charged side-chains, none of the criteria is able to reduce the error rate very significantly. For these amino acids the error rates are 1·4, 1·3 and 1·4 times lower for buried, hydrogen bonded and reproducible side-chains, respectively.

(c) *Using less information*

Encouraged by the accuracy of segment match modeling using all C^α atoms, attempts were made to use less information. This is done in two ways; (1) provide co-ordinates for a subset of the C^α atoms and (2) introduce errors into the C^α co-ordinates.

(i) *Gaps in the chain*

Table 6 shows the r.m.s. deviations of models built with fewer than all C^α atoms. The Table shows the deviations obtained by energy minimization from the X-ray structure. Overall values are given

Table 6

Effect of fewer C^α atoms on r.m.s. deviations, main-chain torsion angles and peptide orientations

	Protein								Overall[a]	
Property	CRN	PTI	CTF	RNS	LYZ	FXN	TLN	APP	Tight	Free
A. r.m.s. deviation										
(i) C^α atoms										
X-ray[b]	0·55	0·54	0·58	0·56	0·48	0·64	0·74	0·59	—	0·61
All[c]	0·11	0·13	0·10	0·15	0·09	0·16	0·12	0·12	0·12	0·68
2nd	0·25	0·62	0·33	0·34	0·32	0·76	0·53	0·63	0·47	0·87
3rd	0·45	0·55	0·34	0·70	0·98	0·80	1·42	0·78	0·75	1·12
4th	1·19	1·90	1·16	2·14	1·57	1·49	1·90	1·54	1·61	1·78
(ii) *Main-chain*										
X-ray	0·57	0·76	0·61	0·59	0·51	0·67	0·78	0·63	—	0·64
All	0·56	0·51	0·29	0·41	0·39	0·44	0·38	0·37	0·42	0·80
2nd	0·47	0·77	0·38	0·62	0·52	0·93	0·70	0·73	0·65	0·96
3rd	0·67	0·68	0·43	0·87	1·06	0·87	1·42	0·88	0·86	1·19
4th	1·25	1·87	1·13	2·21	1·55	1·49	1·90	1·58	1·62	1·76
(iii) *Side-chain*										
X-ray	0·84	0·84	1·01	1·02	1·02	1·09	1·12	0·83	—	0·97
All	1·57	2·43	1·37	2·00	1·59	1·91	1·94	1·42	1·78	1·98
2nd	2·14	2·34	2·00	2·08	2·20	2·46	2·26	2·38	2·24	2·44
3rd	1·73	2·50	1·99	2·37	2·71	2·35	3·25	2·30	2·40	2·61
4th	2·42	3·83	2·54	3·67	3·16	3·13	3·85	3·16	3·22	3·46
(iv) *All atoms*										
X-ray	0·70	0·72	0·81	0·82	0·81	0·90	0·96	0·73	—	0·81
All	1·12	1·73	0·93	1·41	1·15	1·37	1·37	1·00	1·26	1·48
2nd	1·46	1·72	1·36	1·51	1·59	1·84	1·65	1·69	1·60	1·81
3rd	1·25	1·81	1·36	1·75	2·05	1·75	2·47	1·68	1·77	1·99
4th	1·86	2·99	1·88	3·00	2·48	2·43	3·00	2·43	2·51	2·71
B. *Percentage correct*[d]										
(i) χ_1 *torsion angles*										
X-ray[e]	*100*	*91*	*94*	*83*	*95*	*86*	*86*	*98*	*91*	—
All	*92*	*65*	*89*	*61*	*80*	*63*	*62*	*70*	*72*	*72*
2nd	*84*	*65*	*68*	*58*	*65*	*53*	*56*	*66*	*64*	*65*
3rd	*84*	*63*	*62*	*53*	*67*	*52*	*55*	*62*	*62*	*62*
4th	*65*	*52*	*55*	*46*	*50*	*52*	*48*	*59*	*53*	*51*
(ii) χ_1 *torsion angles*										
X-ray	76	77	89	75	95	71	72	79	79	—
All	67	54	78	48	68	48	51	60	59	61
2nd	57	60	78	42	53	42	47	49	53	54
3rd	62	51	67	34	51	52	45	43	50	53
4th	57	37	56	36	34	39	37	41	42	40
(iii) *Peptide units*[f]										
X-ray	100	97	100	99	100	99	98	99	99	—
All	93	88	98	96	95	94	93	96	95	96
2nd	96	86	100	91	87	83	84	90	90	92
3rd	85	76	98	77	76	85	73	79	82	83
4th	72	50	81	43	58	59	59	59	61	68

[a] Results are reported for tight minimization; overall values, averaged over the 8 proteins, are also given for free minimization.

[b] The row titled X-ray gives the r.m.s. deviation values for the structures obtained after a control free minimization of the X-ray co-ordinates.

[c] Gaps are inserted by eliminating some of the C^α atoms from the information used for segment match modeling.

[d] A torsion angle is taken as wrong if its value is further than 30° from the value in the X-ray structure.

[e] The row titled X-ray gives the percentage correct for the structures obtained after a control tight minimization of the X-ray co-ordinates.

[f] A peptide is taken as wrong if the deviation of the peptide oxygen atom is greater than 1·6 Å.

for minimization that is tightly restrained (tight) and unrestrained (free). As fewer C^{α} atoms are used, the models become progressively less accurate. This is particularly true when using every fourth C^{α} atom, where the overall all-atom r.m.s. rises to 2·51 Å, and the main-chain r.m.s. is 1·61 Å. When using every third C^{α} atom, the r.m.s. values are much more like those obtained with all C^{α} atoms (1·77 Å for all-atoms, 0·86 Å for main-chain). Two proteins, CRN and CTF are modeled exceptionally well from every third C^{α} atom with all-atom r.m.s. deviations of 1·25 Å and 1·36 Å, respectively. It is remarkable that such small deviations can be obtained when using the known positions of only 16 and 23 C^{α} atoms for CRN and CTF, respectively.

Unrestrained energy minimization starting at the X-ray structure causes significant r.m.s. deviation of the main chain (0·64 Å), which is comparable with that obtained when modeling from all C^{α} atoms (0·80 Å). The side-chains of the relaxed X-ray structure are much closer to the actual X-ray structure than for the models (r.m.s. of 0·97 Å as opposed to 1·98 Å). The large deviation caused by energy minimization is the reason the known C^{α} atoms are restrained to their X-ray positions. It seems likely that the deviation is caused by inappropriate energy terms for the part of the potential energy function affecting main-chain conformation.

The C^{α} chain paths of models of PTI with every second, third or fourth C^{α} atom are compared in Figure 5 with the path in the X-ray structure. It is interesting that the PTI model built from every third C^{α} atom is closer to the X-ray structure than that built with every second C^{α} atom. The PTI model built using every fourth C^{α} atom has a high C^{α} r.m.s. deviation of 1·90 Å. Two other small proteins, CRN and CTF, did better with every fourth C^{α} atom (Fig. 6) in that the modeled chain path follows the actual path accurately. It is remarkable that building a model based on the known positions of a quarter of the C^{α} atoms works so well.

The models generated using different fractions of the C^{α} atoms had a range of main and side-chain deviations from the X-ray structures. A plot of the side-chain r.m.s. against the main-chain r.m.s. (see Fig. 7) shows a linear relationship that is independent of protein size or number of gaps. Clearly, any improvement of the side-chains will depend on better modeling of the main-chain.

(ii) *Errors in the co-ordinates provided*

Segment match modeling works well when using the exact C^{α} co-ordinates. Tests done with increasingly large errors in positions of the C^{α} atoms (Table 7) show that the modeling accuracy is independent of the magnitude of the error for r.m.s. errors of less than 1 Å. For example, the overall all-atom r.m.s. is 1·66 Å with a 1 Å error and 1·48 Å with no error. Comparison of deviations obtained with fewer C^{α} atoms and with errors shows that a 2 Å error is roughly comparable to using every fourth C^{α} atom, whereas a 1 Å error is comparable to using all C^{α} atoms. The lack of dependence of the accuracy of the model on the accuracy of the co-ordinates indicates that segment match modeling can be used to refine co-ordinates produced by an initial tracing of an electron density map (Jones & Thirup, 1986).

Table 7

Effects of C^{α} errors on r.m.s. deviations of free minimization models

	r.m.s. deviation with C^{α} error[a] (Å)				
Protein	0·0	0·25	0·5	1·0	2·0
A. *C^{α} atoms*					
CRN	0·64	0·63	0·62	0·69	1·65
PTI	0·68	0·69	0·76	0·74	1·57
CTF	0·53	0·59	0·56	0·79	1·58
RNS	0·77	0·65	0·67	0·74	1·76
LYS	0·66	0·64	0·60	0·71	1·44
FXN	0·67	0·76	0·76	0·77	1·53
TLN	0·77	0·72	0·71	0·79	1·60
APP	0·68	0·74	0·74	0·82	1·48
Overall	0·68	0·68	0·68	0·76	1·58
B. *Main-chain atoms*					
CRN	0·92	0·67	0·72	0·93	1·77
PTI	0·81	0·88	0·96	0·90	1·64
CTF	0·59	0·72	0·68	0·87	1·58
RNS	0·86	0·71	0·73	0·87	1·81
LYZ	0·76	0·78	0·75	0·83	1·51
FXN	0·78	0·84	0·88	0·88	1·62
TLN	0·88	0·86	0·87	0·93	1·69
APP	0·76	0·83	0·82	0·90	1·52
Overall	0·80	0·79	0·80	0·89	1·64
C. *Side-chain atoms*					
CRN	1·71	1·77	1·67	2·10	3·00
PTI	2·60	2·42	2·49	2·77	3·64
CTF	1·52	1·74	1·78	2·31	3·45
RNS	2·26	2·03	2·17	2·17	4·19
LYZ	1·86	2·02	2·05	2·05	3·37
FXN	2·07	2·08	2·24	2·20	3·50
TLN	2·17	2·16	1·96	2·20	3·58
APP	1·65	1·79	1·77	2·07	3·11
Overall	1·98	2·00	2·02	2·23	3·48
D. *All atoms*					
CRN	1·33	1·27	1·23	1·56	2·38
PTI	1·91	1·81	1·87	2·04	2·80
CTF	1·10	1·28	1·29	1·67	2·58
RNS	1·68	1·50	1·59	1·62	3·18
LYZ	1·42	1·52	1·54	1·56	2·60
FXN	1·55	1·57	1·68	1·66	2·70
TLN	1·63	1·62	1·50	1·66	2·76
APP	1·25	1·35	1·34	1·54	2·38
Overall	1·48	1·49	1·51	1·66	2·67

Uniformly distributed random errors are added to all the C^{α} co-ordinates used for segment match modeling with all C^{α} atoms. The same errors are used for each of the 10 models as otherwise the averaging procedure would be expected to eliminate all errors. The r.m.s. deviations are presented for models obtained after free minimization in which the C^{α} atoms are able to move from their erroneous starting positions.

[a] C^{α} error is the initial r.m.s. error of the C^{α} atoms.

(d) *Torsion angles*

The accuracy of modeled side-chain conformations can also be judged by the values of the side-chain torsion angles. Table 6 gives the percentage correct χ_1 and χ_2 torsion angles in the models generated here. By way of reference, energy minimization of the X-ray structure causes about 10% of the χ_1 angles and 20% of the χ_2 angles to deviate by more

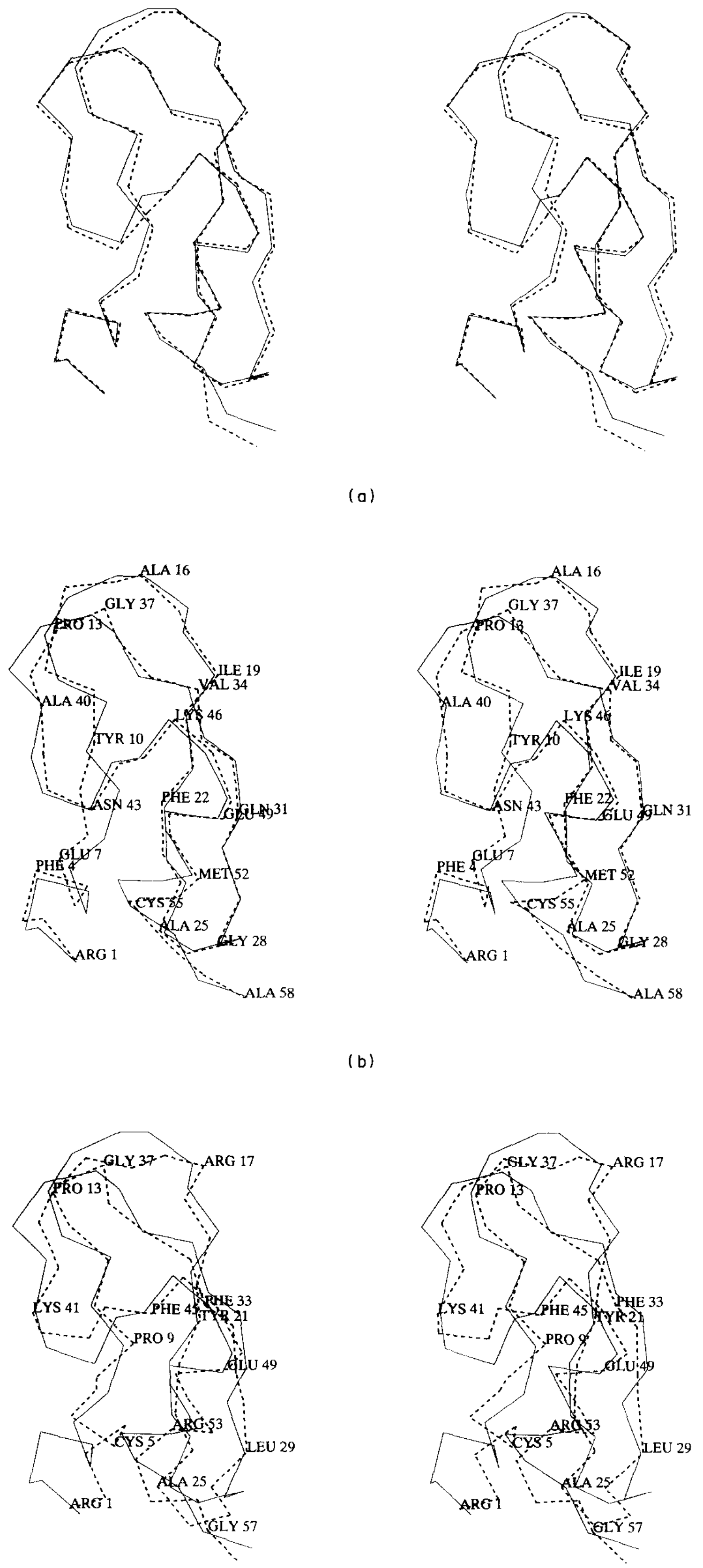


Figure 5. The C^{α} chain path in the pancreatic trypsin inhibitor (PTI) models built using (a) every 2nd C^{α} position, (b) every 3rd C^{α} position and (c) every 4th C^{α} position. The X-ray structure is drawn in continuous lines and the modeled structure is shown in broken lines. In each case, segment match modeling has produced a full set of atomic co-ordinates but only the C^{α} chain is shown for greater clarity. When modeling with 2 or 3 gaps, labels mark the residues for which C^{α}

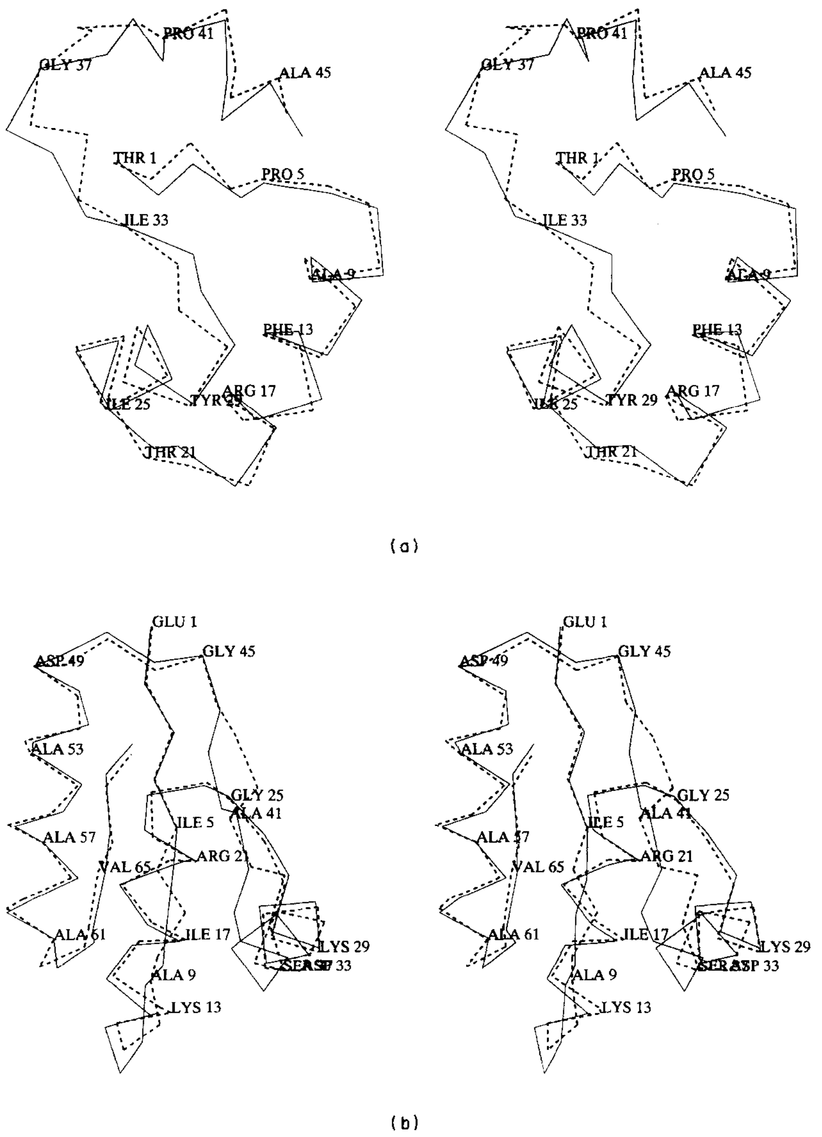


Figure 6. The C^{α} chain path in models of (a) crambin (CRN) and (b) L7/L12 ribosomal protein fragment (CTF) built using every 4th C^{α} position. For both these proteins, the results are much better than for trypsin inhibitor (see Fig. 5) in that the X-ray chain path (continuous line) remains close to the modeled path (broken line); the corresponding r.m.s. deviations for the C^{α} atoms are 1·19 Å and 1·16 Å, respectively. Residues used to provide the C^{α} co-ordinates are labeled.

than 30° from the X-ray values. The overall side-chain r.m.s. deviation of these structures is 0·97 Å, which is considered to be a very small value. The percentage of correct χ_1 angles is also about 10% higher for the modeled structures (see Fig. 8).

The best models, generated using all C^{α} atoms, had between 62% and 92% of χ_1 angles correct (72%, overall) and between 48% and 78% of χ_2 correct (60%, overall). As fewer C^{α} atoms are used, the number of correct angles decreases to 52% for χ_1 and 41% for χ_2 (every 4th). Given that the side-chain r.m.s. deviation is 3·46 Å for this structure, it is clear that the percentage χ torsion angles correct is not a very sensitive criterion. Table 6 gives overall percentages correct for both free and restrained (tight) minimization; the values are very similar. By contrast, the side-chain r.m.s. deviation is always significantly smaller with the tight minimization (see Table 6).

The percentage of correct peptide units is also given in Table 6. The best models have about 95% of the peptide groups in correct conformations, which is much lower than the percentage in the minimized X-ray structures (99%). With larger

co-ordinates are assumed to be known. The model and X-ray chain paths are very close when using every 2nd C^{α} position or every 3rd C^{α} position. When using every 4th C^{α} position, there are greater differences. In particular, there is a rigid body shift of the 3_{10} helix (residues 2 to 7) that is coupled to a distortion of the C terminus of the α-helix (residues 54 to 56). The C^{α} r.m.s. deviations are 0·48 Å, 0·75 Å and 1·61 Å, respectively.

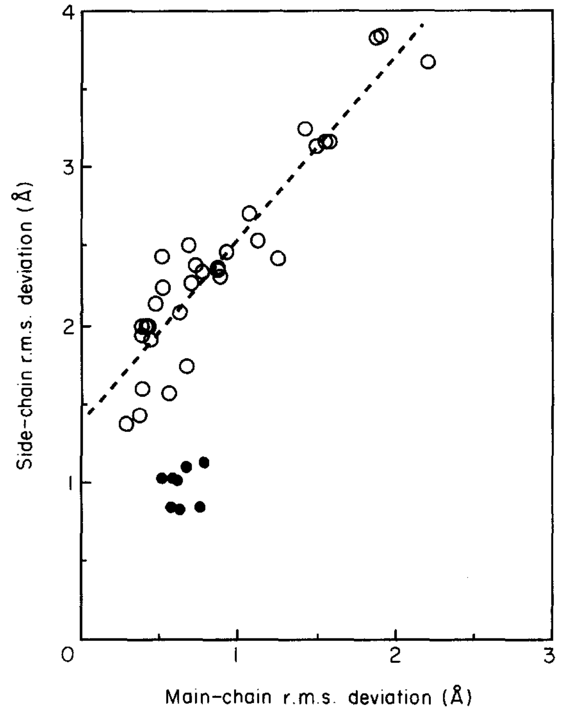


Figure 7. A plot showing the linear relationship between side-chain and main-chain r.m.s. deviations. The values obtained by minimization of the X-ray structures are shown as filled circles. The other 32 values (open circles) represent different segment match modeling using different numbers of C^{α} co-ordinates. The linear relationship between side-chain and main-chain r.m.s. deviation is preserved for different proteins and for different gap sizes.

gaps, the error rate is much higher. Even with models built using every fourth C^{α} atom, 68% of peptide groups are modeled correctly; this is high when one considers that one C^{α} position is specified for four peptide groups.

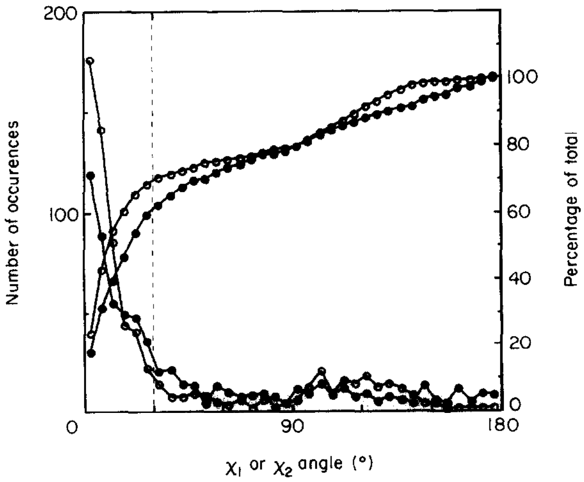


Figure 8. The distribution of χ torsion angle errors for all 973 non-Gly and non-Ala side-chains in the 8 proteins modeled. χ_1 angles (shown as open circles) are modeled more accurately than χ_2 angles (shown as filled circles).

Table 8

Energies of X-ray and modeled structures

Protein	All-atom r.m.s.[d] (Å)	Energy[a] (kcal/mol) Total	/Atom[b]	Diff[c]	Diff (%)
CRN X-ray	0·70	−374	−1·14	0	0·0
All	1·33	−356	−1·07	18	4·8
Every 2nd	1·55	−362	−1·10	12	3·2
Every 3rd	1·41	−353	−1·06	21	5·6
Every 4th	1·65	−353	−1·06	21	5·6
PTI X-ray	0·72	−452	−0·98	0	0·0
All	1·91	−466	−1·02	−14	−3·1
Every 2nd	1·78	−432	−0·94	20	4·4
Every 3rd	1·90	−420	−0·91	32	7·1
Every 4th	3·50	−371	−0·80	81	17·9
CTF X-ray	0·81	−622	−1·26	0	0·0
All	1·10	−621	−1·26	1	0·1
Every 2nd	1·45	−607	−1·24	15	2·4
Every 3rd	1·70	−595	−1·21	27	4·3
Every 4th	2·08	−564	−1·15	58	9·3
RNS X-ray	0·82	−1151	−1·20	0	0·0
All	1·68	−1128	−1·17	23	2·9
Every 2nd	1·85	−1086	−1·14	65	5·6
Every 3rd	2·03	−1046	−1·09	105	9·1
Every 4th	3·27	−907	−0·94	244	21·2
LYZ X-ray	0·81	−1262	−1·21	0	0·0
All	1·55	−1219	−1·18	43	3·4
Every 2nd	1·64	−1192	−1·15	70	5·5
Every 3rd	2·40	−1092	−1·04	170	13·5
Every 4th	2·74	−1043	−1·00	219	17·4
FXN X-ray	0·90	−1225	−1·13	0	0·0
All	1·55	−1287	−1·18	−62	−5·1
Every 2nd	2·08	−1144	−1·05	81	6·6
Every 3rd	1·94	−1236	−1·15	−11	−0·9
Every 4th	2·66	−1086	−1·00	139	11·3
TLN X-ray	0·96	−3267	−1·33	0	0·0
All	1·63	−3196	−1·30	71	2·2
Every 2nd	1·91	−3094	−1·26	173	5·3
Every 3rd	2·60	−3002	−1·22	265	8·1
Every 4th	3·15	−2844	−1·16	423	12·9
APP X-ray	0·73	−3129	−1·31	0	0·0
All	1·25	−3034	−1·27	95	3·0
Every 2nd	1·81	−2812	−1·18	317	10·1
Every 3rd	1·96	−2782	−1·16	347	11·1
Every 4th	2·62	−2601	−1·08	528	16·9

[a] The atomic co-ordinates of the model structures and X-ray structures are both subjected to 600 cycles of energy minimization without any restraints on the positions of the C^{α} atoms. Energy parameters used are those given by Levitt 1983*a*.

[b] /Atom is the total energy divided by the number of non-hydrogen atoms in the particular protein.

[c] Diff is the difference between the total energies of the models and the X-ray structures.

[d] The r.m.s. deviations of all non-hydrogen atoms are relative to the un-minimized X-ray co-ordinates.

(e) *Energy refinement*

Energy minimization is used to refine the mean structures generated by segment match modeling. During this refinement the stereochemistry of the structure is improved as bond lengths, bond angles and torsion angles are standardized, short non-bonded contacts are eliminated and hydrogen bonds formed where possible. The value of the energy after free minimization should indicate the extent of favorable interactions in the models and is analyzed further (see Table 8).

As is to be expected, the total energy of the minimized X-ray co-ordinates depends on the size of

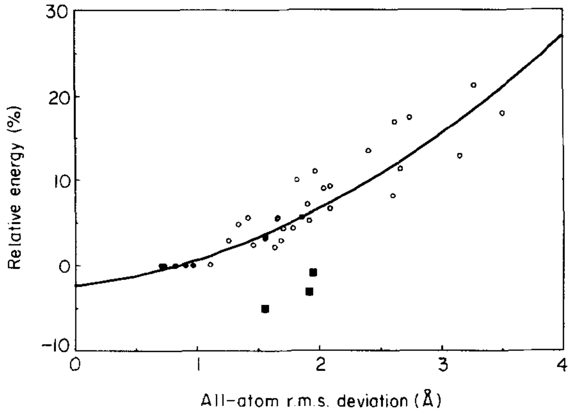


Figure 9. A plot showing the relationship between calculated potential energy and all-atom r.m.s. deviation from the X-ray structure. Results are presented for all the 40 structures generated by 600 steps of unrestrained (free) energy minimization of (a) the 8 X-ray structures (shown as filled circles) and (b) the 4 models generated by segment match modeling using all, every 2nd, every 3rd or every 4th C^α co-ordinate, respectively. The relative energy is expressed as 100 $(U_i - U_i^X) / U_i^X$, where U_i is the total potential energy after minimization of the particular model and U_i^X is the corresponding energy for the minimized X-ray structure. In general, there is a trend for models with higher energy values to deviate more from the X-ray structure. Of the 32 modeled structures, 3 actually have lower energies than the corresponding X-ray structure and are shown as filled squares. The relationship between energy and r.m.s. deviation is best for r.m.s. values above 2 Å, which suggests that energy can be used to discriminate between less accurate models and that further energy minimization could be used to improve such models.

the protein. The energy per atom also increases with protein size but certain molecules are unusual (PTI has only −0·98 kcal/mol atom whereas TLN has −1·33 kcal/mol atom). There seems to be no simple correlation between the ease of modeling and this energy density. The percentage difference in total energy is more well-behaved. The plot of percentage energy difference against r.m.s. deviation (see Fig. 9) presents data for all the 40 minimized structures in Table 8 and shows that the structures with high r.m.s. deviations did have less favorable interactions.

The structures derived by minimization from the X-ray structures generally had lower energies than any of the corresponding models. There are three exceptions: two models built with all C^α atoms (PTI and FXN) and the model of FXN built with every third C^α atom (these are shown as filled circles in Fig. 6). This points out the difficulty of using energy minimization to refine the models further. For large r.m.s. deviations from the X-ray structure (above 2Å) more extensive energy minimization used together with annealing (Levitt, 1983b) may be able to improve the fit to the X-ray structure. For smaller r.m.s. deviations, more minimization may converge on a low-energy structure that is different from the X-ray structure. As it is unlikely that minima with lower free energy really exist near the X-ray structure, this must reflect a defect in the potential energy function.

4. The Effect of Parameters on Segment Match Modeling

Segment match modeling is controlled by a number of different parameters that must be set to specified values. Table 9 shows the effects of the changes in these parameters on the overall main-chain, side-chain and all-atom r.m.s. deviation of the models generated with different fractions of known C^α atoms.

The random number seed has no effect on the r.m.s. deviations of the models; essentially the same results are obtained with five different random numbers. The number of independent models averaged to give the mean model is more important; the r.m.s. deviations are significantly higher when averaging fewer than five models; averaging 20 models gives better results than the standard value of ten used here.

A surprising finding is the relative insensitivity to the number and accuracy of the proteins in the data base. Use of all 76 proteins listed in Table 1 is only slightly better (1·18 Å all-atom r.m.s.) than using the 43 structures with resolution better than 1·8 Å (1·19 Å all-atom r.m.s.) or the 34 remaining lower-resolution structures (1·25 Å all-atom r.m.s.). Use of a minimal data base consisting of seven of the eight test proteins (the protein being modeled is always excluded) also works quite well (1·45 Å all-atom r.m.s.) although now there are only about 1000 residues being used to provide standard conformations.

The segment length, L, which is specified separately for main-chain and side-chain modeling, does affect the r.m.s. deviation, particularly that of the side-chains. For side-chains, increasing the segment length from three to five leads to significantly worse models. It is interesting that the optimum values are $L_m = 4$ for main-chain and $L_s = 3$ for side-chain. Using these L values together gives more accurate models than obtained by our standard conditions of $L = 3$ for both main and side-chain modeling.

How important is it to ensure that the segment from the data base does not clash with the growing target structure? This is tested by not checking the van der Waals' energy. For main-chain, omitting the check had no detectable effect; for side-chains there is a significant effect with the side-chain r.m.s. increasing from 1·65 Å with the check to 2·13 Å without it.

One of the most complicated parts of segment match modeling is the choice of the best fragment using the pseudo-energy composed of r.m.s. deviation and van der Waals' energy (see Methods, section (c)(iv)). This choice is made with some

Table 9
Effect of parameters on overall r.m.s. deviations

Data			Parameters[a]						r.m.s. deviation (Å)			
			Main			Side						
base	N_R	N_{av}	L	β	N_c	L	β	N_c	C$^\alpha$	Main	Side	All
A. *All C$^\alpha$*												
(i) *Vary random number*												
All	**1**	10	3	0	3	3	0	3	—	0·45	1·69	1·21
All	**2**	10	3	0	3	3	0	3	—	0·46	1·69	1·21
All	**3**	10	3	0	3	3	0	3	—	0·46	1·70	1·22
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
All	**5**	10	3	5	3	3	0	3	—	0·45	1·68	1·20
(ii) *Vary number of models averaged*												
All	4	**1**	3	0	3	3	0	3	—	0·52	2·25	1·59
All	4	**2**	3	0	3	3	0	3	—	0·49	1·88	1·34
All	4	**5**	3	0	3	3	0	3	—	0·46	1·68	1·20
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
All	4	**20**	3	0	3	3	0	3	—	0·44	1·61	1·15
(iii) *Vary data base size*												
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
1st	4	10	3	0	3	3	0	3	—	0·45	1·66	1·19
2nd	4	10	3	0	3	3	0	3	—	0·46	1·75	1·25
7pr	4	10	3	0	3	3	0	3	—	0·51	2·04	1·45
(iv) *Vary* L *(zone length)*												
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
All	4	10	**3**	0	3	**4**	0	3	—	0·46	1·87	1·32
All	4	10	**3**	0	3	**5**	0	3	—	0·45	1·97	1·39
All	4	10	**4**	0	3	**3**	0	3	—	0·40	1·62	1·15
All	4	10	**4**	0	3	**4**	0	3	—	0·40	1·80	1·27
All	4	10	**4**	0	3	**5**	0	3	—	0·38	2·03	1·42
All	4	10	**5**	0	3	**3**	0	3	—	0·45	1·65	1·18
All	4	10	**5**	0	3	**4**	0	3	—	0·45	1·80	1·28
All	4	10	**5**	0	3	**5**	0	3	—	0·44	1·93	1·37
(v) *Omit van der Waals' checking*												
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
All	4	10	3	**−1**	3	3	0	3	—	0·44	1·66	1·18
All	4	10	3	**−1**	3	3	**−1**	3	—	0·44	2·13	1·50
(vi) *Vary* N$_c$ *(choice subset) with $\beta = 0$*												
All	4	10	3	0	**1**	3	0	**1**	—	0·55	1·82	1·31
All	4	10	3	0	**2**	3	0	2	—	0·48	1·65	1·19
All	**4**	**10**	**3**	**0**	**3**	**3**	**0**	**3**	—	**0·46**	**1·65**	**1·18**
All	4	10	3	0	**4**	3	0	4	—	0·44	1·69	1·20
All	4	10	3	0	**5**	3	0	5	—	0·43	1·70	1·21
(vii) *Vary β (Boltzmann factor)*												
All	4	10	3	**10**	**5**	3	**10**	**5**	—	0·46	1·95	1·38
All	4	10	3	**·2**	3	3	**2**	3	—	0·43	2·02	1·42
All	4	10	3	**2**	3	3	**2**	3	—	0·44	2·06	1·45
All	4	10	3	**10**	3	3	**10**	3	—	0·49	2·12	1·50
B. *Every 2nd*												
All	**4**	**10**	**5**	**10**	**5**	**5**	**10**	**5**	**0·65**	**0·74**	**2·15**	**1·58**
All	4	**20**	5	10	5	5	10	5	0·63	0·72	2·14	1·57
All	4	10	5	10	5	3	10	5	0·66	0·74	2·24	1·63
All	**4**	**10**	**5**	**0**	**3**	**3**	**0**	**3**	**0·67**	**0·77**	**2·13**	**1·56**
All	4	10	5	0	**2**	3	0	2	0·70	0·79	2·13	1·57
C. *Every 3rd*												
All	**4**	**10**	**5**	**10**	**5**	**5**	**10**	**5**	**1·03**	**1·07**	**2·42**	**1·84**
All	4	**20**	5	10	5	5	10	5	0·98	1·02	2·35	1·78
All	4	10	5	10	5	3	10	5	1·02	1·05	2·55	1·92
All	**4**	**10**	**5**	**0**	**3**	**3**	**0**	**3**	**1·10**	**1·15**	**2·66**	**2·01**
All	4	10	5	0	**2**	3	**0**	2	1·10	1·14	2·68	2·02

Table 9 *(Continued)*

Data base			Parameters[a] Main			Side			r.m.s. deviation (Å)			
	N_R	N_{av}	L	β	N_c	L	β	N_c	C^α	Main		
D. *Every 4th*												
All	**4**	**10**	**5**	**10**	**5**	**5**	**10**	**5**	**1·79**	**1·77**	**3·24**	**2·57**
All	4	**20**	5	10	5	5	10	5	1·66	1·64	3·06	2·42
All	4	10	5	10	5	**3**	10	5	1·69	1·69	3·32	2·59
All	**4**	**10**	**5**	**0**	**3**	**3**	**0**	**3**	**1·81**	**1·81**	**3·47**	**2·72**
All	4	10	5	**0**	**2**	**3**	**0**	**2**	1·85	1·84	3·50	2·75

The r.m.s. deviations are the overall values averaged over the 8 test proteins using the mean structures before any energy minimization. Such minimization will increase these deviation values slightly (see Table 2).

[a] The parameters are as follows: Data base is **All**, for all 76 proteins, **1st**, for 43 high-resolution proteins, **2nd**, for 34 lower resolution proteins and **7pr**, for 7 of the 8 test proteins. N_R is the random number seed. N_{av} is the number of independent models averaged. L is the segment length. β is the Boltzmann weight. N_c is the number of top matches selected from (these last 3 parameters can be specified independently for main-chain and side-chain modeling). Bold-face indicates standard parameter values and the parameters that have been changed.

degree of randomness controlled by the parameter β, which acts like an inverse temperature. When β is high, the best segment is more likely to be chosen; when β is zero, any segment is chosen. This choice is further controlled by N_c, the number of good segments from which the segment is chosen. The standard conditions set $\beta = 0$ and $N_c = 3$. Good results are obtained for N_c between 2 and 5 (see Table 9).

When one considers models built with fewer known atoms, the role of positive β is more important. The standard conditions have to be amended in that the main-chain segment length, L, must be increased from three to five to prevent instabilities in the co-ordinate fitting procedures. When using every third or fourth C^α atom, significantly better results are obtained with $\beta = 10$ and $N_c = 5$. To accommodate this situation, the program has two sets of standard conditions. When there are no missing C^α atoms in the segment, the parameters are $L = 3$ for main and side-chain, $\beta = 0$ and $N_c = 3$. When there are missing C^α atoms in the segment, the parameters are $L_m = 5$ for main-chain, $L_s = 3$ for side-chain, $\beta = 10$ and $N_c = 5$. This choice is made completely automatically for each segment under construction so that no human intervention is required.

5. Discussion

(a) *How good are the models?*

When using all C^α atoms, the present method is able to calculate positions for main-chain atoms with an accuracy of 0·4 Å and side-chain atoms with an accuracy of 1·8 Å. How good are these values? Judging the significance of r.m.s. deviation values is subjective. It is further complicated by the very non-linear nature of this measure: an r.m.s. deviation of 3 Å is much more than twice as bad as an r.m.s. deviation of 1·5 Å. On the other hand, the accuracy of the calculated models is crucial if one is to be able to use these methods to help refine X-ray structure, model homologous proteins, etc.

The most direct criterion is visual inspection of modeled and X-ray co-ordinates superimposed in stereo drawings. For example, in Figure 3(a), it is clear that 2 of the 11 aromatic side-chain are incorrect (side-chain r.m.s. deviation great than 2 Å). When using all C^α atoms, almost 90% of the side-chains are modeled with this accuracy. When using every second C^α atom, the overall side-chain r.m.s. increases from 1·78 Å to 2·24 Å and the number of correctly modeled side-chains drops to 75%. Comparison of Figure 3(a) and (b) clearly shows that the prediction of the aromatic side-chains of lysozyme gets worse when using half the C^α positions.

There are two computationally derived measures of r.m.s. deviation that can help appreciate the accuracy of the present method, energy minimization and molecular dynamics. Energy minimization from the X-ray structure is always found to cause a shift of atomic co-ordinates. The extent of this deviation depends on strength of the restraint to the X-ray structure. With no restraint, the r.m.s. deviation is 0·64 Å for the main-chain and 0·97 Å for the side-chains. The main-chain value is much larger than that obtained by segment match modeling using a tight restraint (0·42 Å, see Table 6) and comparable with that obtained using every second C^α atom (0·66 Å). This indicates that energy minimization is spoiling the fit of the backbone atoms to the X-ray data. This must be due to a defect in the energy parameters that had not been apparent till now. Until the energy functions are corrected, it will not be possible to model the main chain more accurately than about 0·5 Å; this value is small enough for most purposes and much smaller than the deviations of the side-chains. The side-chain deviations obtained by free minimization of the X-ray structures are much smaller than those obtained by modeling (0·97 Å compared with 1·78 Å).

Molecular dynamics that also starts from the X-ray structure is more powerful than energy minimization in two respects, the conformation can move further than the first local energy minimum and solvent can be included. In the relatively accurate simulations of the small protein BPTI in solution (Levitt & Sharon, 1988), the main-chain r.m.s. deviation from the X-ray structure reaches a plateau value of 1·2 Å, while the corresponding deviation of the side-chain atoms reaches 1·50 Å. These values are comparable with those found for BPTI modeled from either all or only half the C^{α} atoms.

There is an experimental measure that also helps appreciate the accuracy of the modeling: the r.m.s. deviation between protein co-ordinates before and after refinement. Protein co-ordinates at different stages of refinement can have missing atoms (usually only the side-chains) or duplicated atoms (multiple side-chain conformations in very well-refined structures) making calculation of the r.m.s. deviation problematic. Restricting attention to the six cases where the two proteins have the same number of atoms and the C^{α} r.m.s. deviation is less than 0·6 Å, gives r.m.s. differences of 0·16 to 0·67 Å for main-chain and of 0·28 Å to 1·14 Å for side-chain. Segment match modeling gives very similar r.m.s. deviations (0·40 Å from main chain, 1·28 Å for all-atoms).

(b) *Comparison with other methods*

There have been a number of other attempts to model protein co-ordinates using the amino acid sequence and some known atomic co-ordinates. In comparing these methods with segment match modeling attention is focused on three aspects: (1) is the modeling automatic or manual; (2) what atoms are modeled (main-chain only or main-chain and side-chain); (3) how much information is used (all C^{α} atoms, all main-chain atoms, main-chain and C^{β} atoms). In comparing the success of other methods with SMM, two criteria are used; (1) the accuracy of the modeled co-ordinates (r.m.s. deviation from the known X-ray structure) and (2) the dependence on known atomic positions.

Jones & Thirup (1986) pioneered the use of a data base of segments from known proteins. While their method, described more fully by Jones *et al.* (1991), has proved invaluable in the interpretation of electron density maps, it is not automatic and has not been systematically tested in modeling from partial co-ordinate data.

Another manual modeling attempt involved building one protein, flavodoxin, from the known C^{α} co-ordinates (Reid & Thornton, 1989). Although their structure is not generated automatically, these experienced researchers expended much effort to ensure that the modeling was objective and based on well-defined rules. The r.m.s. deviation obtained for the main-chain atoms, which were generated by the method of Jones & Thirup (1986), is 0·51 Å, whereas that obtained for the side-chain atoms is 2·41 Å. These values have served as standards for several of the automatic schemes. In the present work, the corresponding values of 0·44 Å and 1·91 Å, respectively (see Table 2), are significantly better. In fact, the model of flavodoxin built here with every third C^{α} atom has r.m.s. deviations of 0·80 Å for main-chain and 2·35 Å for side-chain atoms (Table 6). These values are comparable with the value obtained by Reid & Thornton (1989) using three times as many known C^{α} positions.

The first study to use automatic search of a segment data base (Claessens *et al.*, 1989) focused on building the main chains of three proteins (triose phosphate isomerase, citrate synthase and carboxypeptidase) from the positions of all the C^{α} atoms. The average r.m.s. deviation they obtain for the main chain is 0·61 Å, which is higher than the overall value of (0·42 Å, see Table 2) found here for eight different proteins. In fact, their value obtained using every C^{α} atom is more comparable with our value obtained using every second C^{α} atom (0·66 Å, see Table 6). They did, however, point out that the rebuilt backbone is relatively insensitive to errors in the C^{α} positions. This insensitivity, which was also observed by Holm & Sander (1991) and in the present results may be a property of the data base approach.

Molecular dynamics was used to build protein structures from the C^{α} co-ordinates in an intriguing study that did not use any data base of known protein structures (Correa, 1990). Results for three proteins (α-lytic protease, troponin C^{α} and flavodoxin) are impressive with main-chain r.m.s. deviations ranging from 0·19 Å to 0·49 Å, and all-atom r.m.s. deviations from 1·24 Å to 1·68 Å. His results for flavodoxin of 0·49 Å for main chain and 1·64 Å for all-atoms are worse than the values obtained here (0·44 Å and 1·37 Å, respectively, see Table 2). The method requires considerable computer resources: modeling a protein with 200 residues takes about 40 hours on the Star ST-50 computer using a highly optimized version of the program GEMM. This is equivalent to at least 100 hours on the Silicon Graphics 4D/25. Nevertheless, it is very impressive that molecular dynamics can produce such accurate models without recourse to a data base of refined structures. These results suggest that extensive molecular dynamics followed by energy minimization is able to reproduce the details of the native protein conformation.

There have been three recently published studies automatically modeling side-chain co-ordinates from a known main-chain structure. The first of these (Lee & Subbiah, 1991) is unique in that it does not make any use of a data base of segments or a library of side-chain conformations. Instead, simulated annealing is used to optimize the packing of side-chains using van der Waals' interactions. Results obtained on the same sample of eight test proteins used here give overall side-chain r.m.s. deviations of 1·77 Å, which is very similar to the value of 1·78 Å obtained here. These results are most impressive especially if one considers that their r.m.s. value omits Ala and Pro side-chains and C^{β} atoms (the deviation of these atoms is lower than average, so that a comparison over the same sets of atoms would favor Lee & Subbiah). The present study requires fewer known atomic positions (the set of all C^α atoms is 0·2 the size of the set of main-chain plus C^β atoms) and runs about 20 times faster (30 min for FXN as opposed to 600 min). The Lee & Subbiah (1991) method does not rely on a data base of known protein structures making its excellent performance very impressive. The major drawback to this approach, in contrast to the present method, is that the main-chain atoms are required and the main chain is held fixed.

Holm & Sander (1991) have used a method that combines a data base search for the main-chain conformation with simulated annealing for the side-chains. When tested on the C^α co-ordinates of known proteins, the r.m.s. deviation of the main chain is between 0·4 Å and 0·6 Å. This is comparable with the overall main-chain r.m.s. deviation of 0·42 Å found here. The r.m.s. of the side-chains added to these models have r.m.s. deviations of 2·21 Å, compared to the value of 1·78 Å found here. The percentage of correct χ_1 angles is slightly lower than found here (70% as against 72%, see Table 6). Their method is fast, allowing a side-chain to be built in four seconds. Tuffery *et al.* (1991), who use a method similar to that used by Holm & Sander (1991), get similar results in a test that assumes that all backbone atoms are known (side-chain r.m.s. of 1·84 Å and 1·91 Å, respectively).

Overall, the segment match modeling procedure is more powerful than all other published methods in that it works with less information and still produces the most accurate models. For example, the overall r.m.s. deviation of side-chain atoms found here is as good when using only half the C^α atoms (2·26 Å) as Holm and Sander (1991) achieve using all C^α atoms (2·21 Å). The accuracy achieved here using all C^α atoms (1·78 Å) is better than that achieved by Holm & Sander (1991) or Tuffery *et al.* (1991) using all the backbone atoms (1·91 Å and 1·84 Å, respectively). These overall r.m.s. deviation values are for different sets of proteins so the r.m.s. deviation values are only indicative of the general accuracy. The present method is also insensitive to errors in the C^α co-ordinates of up to 1 Å r.m.s.; by contrast, Holm & Sander's method fails when this error exceeds 0·4 Å.

It is not clear why segment match modeling works as well as or better than more complicated search methods that employ combinatorial approaches (Lee & Subbiah, 1991) or simulated annealing (Holm & Sander, 1991; Tuffery *et al.*, 1991). Table 9 shows that two factors are responsible for the success of the simpler segment match modeling method: the averaging of independent models and the use of van der Waals' contact checking. Without both techniques, the overall r.m.s. deviation of side-chains would be in excess of 2·5 Å, so that few side-chains would be correct with r.m.s. deviations below 2 Å. Averaging the co-ordinates of side-chains that have adopted different conformations leads to severely distorted average structures. The distorted side-chain is usually smaller than normal and energy minimization using ENCAD always leads to normal, unstrained stereochemistry. This powerful combination of robust energy minimization and co-ordinate averaging also makes segment match modeling insensitive to junctions between main-chain segment.

(c) *Other applications and improvements*

A general, fully automated method capable of filling in missing atomic co-ordinates as accurately as is achieved with segment match modeling has other applications, including homology modeling and fleshing-out simple folding models with detailed atomic structure.

Homology modeling is a simple extension of segment match modeling from C^α co-ordinates. Instead of adding the native side-chains to the known native C^α framework, one adds the side-chains of the related homologous protein. Gaps in the framework will be of two kinds: (1) insertions where the homologous protein is longer; these additional residues have to be modeled without guide C^α co-ordinates, a situation that resembles test modeling with gaps; and (2) deletions where the homologous protein is shorter; these residues are simply left out and the gap in the chain closed when energy minimization imposes good internal geometry. Segment match modeling has been used to build models of a large number of different homologous proteins (unpublished results). While the models seem reasonable and have helped the experimental study of the particular protein, the method has not been tested systematically. Such a systematic study of segment match modeling on pairs of homologous proteins is now underway.

Segment match modeling can flesh out any C^α chain path. Because the method can work with very little information (for example, from every 4th C^α atom), it can be used to take what is little more than a three-dimensional cartoon sketch of a polypeptide chain fold and convert it to a full set of stereochemically correct co-ordinates. These co-ordinates can then be subjected to further refinement using other modeling methods like simulated annealing to improve side-chain packing (Lee & Subbiah, 1991) and annealing molecular dynamics (Levitt, 1983*b*) to reduce the potential energy and hopefully become more native-like. The speed and robustness of segment match modeling make it ideal for building all-atom models from simple lattice chain paths like those used in low-resolution simulation of protein folding (Levitt & Warshel, 1975; Skolnick & Kolinsky, 1990; Hinds & Levitt, 1992).

Segment match modeling as implemented here makes a number of simplifying assumptions. One possible short-coming is that side-chains are added to the structure one residue at a time and without considering correlation between side-chain packings. This is easiest to understand in the context of disulfide bridges. The present version chooses a conformation of each half of the SS bridge independently using only local information: no use is made of the strong requirement that the two S^γ atoms must be close together. It would be better to treat the bridge as a unit and match both ends of the bridge (as done in our modeling of disulfide bridges in T4 lysozyme, see Matsumura *et al.*, 1989). The same method could be generalized and used for pairs of hydrophobic residues suspected to interact strongly.

Another problem, highlighted by the data in Table 6, is that the r.m.s. deviation increases rapidly when the gap is three residues or longer. Homology modeling will sometimes involve inserting segments that are longer than this and new methods will have to be developed. There are a number of ways to proceed. (1) The segment match modeling procedure could be extended to treat large numbers of gaps (large insertions) by building these regions last (so that the rest of the structure can provide a reliable framework) or by building them more than ten times so as to improve the signal-to-noise ratio. (2) Many possible conformations for the inserted region could be generated by systematic search (Moult & James, 1986) or by molecular dynamics at high temperature (Fine *et al.*, 1986).

Energy functions used to impose good stereochemistry on the modeled structures are not in perfect agreement with the details of the X-ray structure: minimization causes small but significant shifts from that structure (<1 Å). This means that with these energy functions, the lowest energy structures deviate from the X-ray structure. On the other hand, when the r.m.s. deviation from the X-ray structure exceeds 2 Å, the energy is significantly increased (see Fig. 9). The results of Correa (1990) show that extensive molecular dynamics and energy minimization can give stable all-atom structures that are close to the X-ray structure. If energy functions could be improved to fit native structures, these methods could be more powerful in generating correct conformations.

Segment match modeling and the other methods do better for core residues. More specifically, modeling is found to work best for the non-polar residues and worst for the charged residues. Even those polar side-chains that are modeled reproducibly (with low variance) are often inaccurate. This difference in the behavior of non-polar and polar side-chains is puzzling. It may be due to: (1) inadequate treatment of electrostatic interactions; (2) omission of any solvent interactions; or (3) real variability of polar side-chains caused by crystal packing interactions. We are trying to distinguish between these possibilities in the hope of further improving the accuracy of the modeling.

6. Conclusion

Segment match modeling provides a simple yet accurate method for modeling protein conformations from a small number of known atomic positions. The method is accurate (all-atom r.m.s. deviation of better than 1·3 Å), efficient (14 s/residue on an inexpensive workstation), insensitive to random errors (up to 1 Å in C^α positions), and robust (good models can be built using every 3rd C^α position). By always building a number of independent models (normally 10), the method provides the coordinate variance of every atom; the accuracy of the atoms with the smallest standard deviations (more reproducible) is always higher. Overall, segment match modeling out-performs all other methods in terms of its higher accuracy and its lack of dependence on known atomic positions; it is also the only method that provides an estimate of the errors in the modeling.

The success of segment match modeling and other methods for modeling from C^α atoms suggests that much of the detailed information in a protein conformation is redundant. Those methods that use a data base of refined protein X-ray structures (this work; Jones & Thirup, 1986; Claessens *et al.*, 1989; Holm & Sander, 1991; Tuffery *et al.*, 1991) show that conformational themes recur with very high frequency. In this regard it is interesting that a data base of only 1000 amino acid residues works fairly well, yielding an all-atom r.m.s. deviation of about 1·4 Å (see Table 9). The other methods that rely on energy criteria without recourse to any data base (Lee & Subbiah, 1991; Correa, 1990) show that the positions of the C^α atoms are generally sufficient to define side-chain conformations uniquely. Both these results are unexpected; they imply that predictions of the folded structure from the amino acid sequence may also be easier than expected in that the effective number of degrees of freedom may be quite small.

This work was supported by NIH award GM-39552.

References

Abad-Zapatero, C., Griffith, J. P., Sussman, L. & Rossmann, M. G. (1987). Refined crystal structure of dogfish m4 apo-lactate dehydrogenase. *J. Mol. Biol.* **198**, 445–467.

Acharya, K. R., Stuart, D. I., Walker, N. P. C., Lewis, M. & Phillips, D. C. (1989). Refined structure of baboon alpha-lactalbumin at 1·7 Å resolution. Comparison with C-type lysozyme. *J. Mol. Biol.* **208**, 99–127.

Almassy, R. J., Fontecilla-Camps, J. C., Suddath, F. L. & Bugg, C. E. (1983). Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus* Ewing, refined at 1·8 Å resolution. *J. Mol. Biol.* **170**, 497–527.

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.

Arutyunyan, E. G., Kuranova, I. P., Vainshtein, B. K. & Steigemann, W. (1980). X-ray structural investigation of leghemoglobin. VI. Structure of acetate-ferri-leghemoglobin at a resolution of 2·0 Å. *Kristallografiya*, **25**, 80–103.

Baker, E. N. (1988). Structure of azurin from *Alcaligenes denitrificans*. Refinement at 1·8 Å resolution and comparison of the two crystallographically independent molecules. *J. Mol. Biol.* **203**, 1071–1095.

Baker, E. N. & Dodson, E. J. (1980). Crystallographic refinement of the structure of actinidin at 1·7 Å resolution by fast Fourier least-squares methods. *Acta Crystallogr. sect. A*, **36**, 559–572.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rogers, 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.

Betzel, C., Pal, G. P. & Saenger, W. (1988). Synchrotron X-ray data collection and restrained least-squares refinement of the crystal structure of proteinase K at 1·5 Å resolution. *Acta Crystallogr. sect. B*, **44**, 163–172.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rerat, B. & Rerat, C. (1978). Structure of prealbumin, secondary, tertiary and quaternary interactions determined by Fourier refinement at 1·8 Å. *J. Mol. Biol.* **121**, 339–356.

Blevins, R. A. & Tulinsky, A. (1985). The refinement and the structure of the dimer of alpha-chymotrypsin at 1·67-Å resolution. *J. Biol. Chem.* **260**, 4264–4275.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C.-W. (1981). X-ray analysis (1·4-Å resolution) of avian pancreatic polypeptide. Small globular protein hormone. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4175–4179.

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.

Bode, W., Epp, O., Huber, R., Laskowski, M., Jr & Ardelt, W. (1985). The crystal and molecular structure of the third domain of silver pheasant ovomucoid OMSVP3. *Eur. J. Biochem.* **147**, 387–395.

Bode, W., Papamokos, E., Musil, D., Seemueller, U. & Fritz, H. (1986). Refined 1·2 Å crystal structure of the complex formed between subtilisin Carlsberg and the inhibitor eglin c. Molecular structure of eglin and its detailed interaction with subtilisin. *EMBO J.* **5**, 813–818.

Borkakoti, 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.

Carter, C. W., Kraut, J., Freer, S. T., Xuong, N.-H., Alden, R. A. & Bartsch, R. G. (1974). 2-Å crystal structure of oxidized chromatium high potential iron protein. *J. Biol. Chem.* **249**, 4212–4225.

Chambers, J. L. & Stroud, R. M. (1977). Difference-Fourier refinement of the structure of DIP-trypsin at 1·5 Å with a minicomputer technique. *Acta Crystallogr. sect. B*, **33**, 1824–1837.

Claessens, M., Van Cutsem, E., Lasters, I. & Wodak, S. (1989). Modeling the polypeptide backbone with "spare parts" from known protein structures. *Protein Eng.* **2**, 335–345.

Correa, P. E. (1990). The building of protein structures from α-carbon coordinates. *Proteins*, **7**, 366–377.

Cotton, F. A., Hazen, E. E., Jr & Legg, M. J. (1979). Staphylococcal nuclease. Proposed mechanism of action based on structure of enzyme-thymidine 3,5-biphosphate-calcium ion complex at 1·5-Å resolution. *Proc. Nat. Acad. Sci., U.S.A.* **76**, 2551–2555.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). Structure of bovine pancreatic phospholipase A2 at 1·7 Å resolution. *J. Mol. Biol.* **147**, 97–123.

Epp, O., Ladenstein, R. & Wendel, A. (1983). The refined structure of the selenoenzyme glutathione peroxidase at 0·2-nm resolution. *Eur. J. Biochem.* **133**, 51–69.

Eriksson, A. E., Kylsten, P. M., Jones, T. A. & Liljas, A. (1988). Crystallographic studies of inhibitor binding sites in human carbonic anhydrase II. A penta-coordinated binding of the SCN$^-$ ion to the zinc at high pH. *Proteins. Struct. Funct.* **4**, 283–293.

Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1·74 Å resolution. *J. Mol. Biol.* **175**, 159–174.

Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L. & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins*, **1**, 342–362.

Finzel, B. C., Weber, P. C., Hardman, K. D. & Salemme, F. R. (1985). Structure of ferricytochrome C′ from *Rhodospirillum molischianum* at 1·67 Å resolution. *J. Mol. Biol.* **186**, 627–643.

Finzel, B. C., Clancy, L. L., Holland, D. R., Muchmore, S. W., Watenpaugh, K. D. & Einspahr, H. M. (1989*a*). Crystal structure of recombinant human interleukin-1 *β* at 2·0 Å resolution. *J. Mol. Biol.* **209**, 779–791.

Finzel, B. C., Kimatian, S., Ohlendorf, D. H., Wendoloskii, J. J., Levitt, M. & Salemme, F. R. (1989*b*). Molecular modeling with substructure libraries derived from known protein structures. In *Crystallographic and Modeling Methods in Molecular Design* (Bugg, C. E. & Ealick, S. E., eds), pp. 175–188, Springer-Verlag, Berlin.

Frey, M., Sieker, L., Payan, F., Haser, R., Bruschi, M., Pepe, G. & LeGall, J. (1987). Rubredoxin from *Desulfovibrio gigas*. A molecular model of the oxidized form at 1·4 Å resolution. *J. Mol. Biol.* **197**, 525–541.

Fujinaga, M. & James, M. N. G. (1987). Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1·8 Å resolution. *J. Mol. Biol.* **195**, 373–396.

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

Furey, W., Jr, Wang, B. C., Yoo, C. S. & Sax, M. (1983). Structure of a novel Bence-Jones protein (rhe) fragment at 1·6 Å resolution. *J. Mol. Biol.* **167**, 661–692.

Gelin, B. R. & Karplus, M. (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.

Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1·6 Å resolution. *J. Mol. Biol.* **169**, 521–563.

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.

Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N. & Kakudo, M. (1984). Refined structure of cytochrome c3 at 1·8 Å resolution. *J. Mol. Biol.* **172**, 109–139.

Hinds, D. H. & Levitt, M. (1992). A lattice model for protein structure prediction at low resolution. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 2536–2540.

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.

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. (1978). Graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11**, 268–272.

Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J.-Y. & Cowan, S.W. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. sect. A*, **47**, 110–119.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. sect. A*, **34**, 827–828.

Kamphuis, I. G., Kalk, K. H., Swarte, M. B. A. & Drenth, J. (1984). Structure of papain refined at 1·65 Å resolution. *J. Mol. Biol.* **179**, 233–256.

Karplus, P. A. & Schulz, G. E. (1987). Refined structure of glutathione reductase at 1·54 Å resolution. *J. Mol. Biol.* **195**, 701–729.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. *Nature (London)*, **185**, 422–427.

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 Å. *J. Mol. Biol.* **195**, 555–579.

Levitt, M. (1983*a*). Molecular dynamics of native protein. I. 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.

Levitt, M. & Sharon, R. (1988). Accurate simulation of protein dynamics in solution. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 7557–7561.

Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature (London)*, **253**, 694–698.

Lindqvist, Y. (1989). Refined structure of spinach glycolate oxidase at 2 Å resolution. *J. Mol. Biol.* **209**, 151–166.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3·0 Å and 1·9 Å resolution. *J. Mol. Biol.* **141**, 369–391.

Mathews, F. S., Argos, P. & Levine, M. (1972). The structure of cytochrome *b*5 at 2·0 Å resolution. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 387–395.

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.

Matsuura, Y., Takano, T. & Dickerson, R. E. (1982). Structure of cytochrome *c*551 from *Pseudomonas aeruginosa* refined at 1·6 Å resolution and comparison of the two redox forms. *J. Mol. Biol.* **156**, 389–409.

Matthews, D. A., Bolin, J. T., Burridge, J. M., Filman, D. J., Volz, K. W., Kaufman, B. T., Beddell, C. R., Champness, J. N., Stammers, D. K. & Kraut, J. (1985). Refined crystal structures of *Escherichia coli* and chicken liver dihydrofolate reductase containing bound trimethoprim. *J. Biol. Chem.* **260**, 381–391.

McPhalen, C. A. & James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry*, **26**, 261–269.

Meyer, E., Cole, G., Radahakrishnan, R. & Epp, O. (1988). Structure of native porcine pancreatic elastase at 1·65 Å resolution. *Acta Crystallogr. sect. B*, **44**, 26–38.

Miller, M., Jaskolski, M., Rao, J.K.M., Leis, J. & Wlodawer, A. (1989). Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature (London)*, **337**, 576–579.

Moews, P. C. & Kretsinger, R. H. (1975). Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *J. Mol. Biol.* **91**, 201–225.

Mondragon, A., Subbiah, S., Almo, S. C., Drottar, M. & Harrison, S. C. (1989). Structure of the amino-terminal domain of phage 434 repressor at 2·0 Å resolution. *J. Mol. Biol.* **205**, 189–200.

Morize, I., Surcouf, E., Vaney, M. C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E. & Mornon, J. P. (1987). Refinement of the $C222_1$ crystal form of oxidized uteroglobin at 1·34 Å resolution. *J. Mol. Biol.* **194**, 725–739.

Moult, J. & James, M. N. G. (1986). An algorithm for determining the confirmation of polypeptide segments in proteins by systematic search. *Proteins*, **1**, 146–163.

Navia, M. A., McKeever, B. M., Springer, J. P., Lin, T. Y., Williams, H. R., Fluder, E. M., Dorn, C. P. & Hoogsteen, K. (1989). Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1·84-Å resolution. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 7–11.

Petratos, K., Banner, D. W., Beppu, T., Wilson, K. S. & Tsernoglou, D. (1987). The crystal structure of pseudoazurin from *Alcaligenes faecalis* s-6 determined at 2·9 Å resolution. *FEBS Letters*, **218**, 209–214.

Pflugrath, J. W., Wiegand, G., Huber, R. & Vertesy, L. (1986). Crystal structure determination, refinement and the molecular model of the alpha-amylase inhibitor Hoe–467a. *J. Mol. Biol.* **189**, 383–386.

Phillips, D. C. (1966). The three-dimensional structure of an enzyme molecule. *Sci. Amer.* **215**, 78–90.

Phillips, S. E. V. (1980). Structure and refinement of oxymyoglobin at 1·6 Å resolution. *J. Mol. Biol.* **142**, 531–554.

Pletnev, V., Kuzin, A., Trakhanov, S., Popovich, V. & Tsigannik, I. (1982). X-ray investigation of three-dimensional structure of actinoxanthin. In *Chemistry of Peptides and Proteins* (Voelter, W. *et al.*, eds), Walter de Gruyter & Co., Berlin.

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.

Poulos, T. L., Finzel, B. C. & Howard, A. J. (1987). High-resolution crystal structure of cytochrome P450CAM. *J. Mol. Biol.* **195**, 687–700.

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 Å resolution. *Biochemistry*, **22**, 4420–4433.

Reeke, G. N., Jr, Becker, J. W. & Edelman, G. M. (1975). The covalent and three-dimensional structure of concanavalin A. IV. Atomic coordinates, hydrogen bonding, and quaternary structure *J. Biol. Chem.* **250**, 1525–1547.

Rees, D. C., Lewis, M. & Lipscomb, W. N. (1983). Refined crystal structure of carboxypeptidase A at 1·54 Å resolution. *J. Mol. Biol.* **168**, 367–387.

Reid, L. S. & Thornton, J. M. (1989). Rebuilding flavodoxin from C^a coordinates: A test study. *Proteins*, **5**, 170–182.

Remington, S. J., Woodbury, R. G., Reynolds, R. A., Matthews, B. W. & Neurath, H. (1988). The structure of rat mast cell protease II at 1·9-Å resolution. *Biochemistry*, **27**, 8097–8105.

Satyshur, K. A., Rao, S. T., Pyzalska, D., Drendel, W., Greaser, M. & Sundaralingam, M. (1988). Refined structure of chicken skeletal muscle troponin C in the two-calcium state at 2-Å resolution. *J. Biol. Chem.* **263**, 1628–1647.

Saul, F. A., Amzel, L. M. & Poljak, R. J. (1978). Preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin NEW at 2·0 Å resolution. *J. Biol. Chem.* **253**, 585–597.

Skolnick, J. & Kolinski, A. (1990). Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key β-barrel proteins. *J. Mol. Biol.* **212**, 787–817.

Sheriff, S., Hendrickson, W. A. & Smith, J. L. (1987). Structure of myohemerythrin in the azido-met state at 1·7/1·3 Å resolution. *J. Mol. Biol.* **197**, 273–296.

Sielecki, A. R, Hendrickson, W. A., Broughton, C. G., Delbaere, L. T. J., Brayer, G. D. & James, M. N. G. (1979). Protein structure refinement. *Streptomyces griseus* serine protease A at 1·8 Å resolution. *J. Mol. Biol.* **134**, 781–804.

Smith, J. L., Corfield, P. W. R., Hendrickson, W. A. & Low, B. W. (1988). Refinement at 1·4 Å resolution of a model of erabutoxin B. Treatment of ordered solvent and discrete disorder. *Acta Crystallogr. sect. A*, **44**, 357–368.

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.* **117**, 195–225.

Steigemann, W. & Weber, E. (1979). Structure of erythrocruorin in different ligand states refined at 1·4 Å resolution. *J. Mol. Biol.* **127**, 309–338.

Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1983). Adjustment of restraints in the refinement of methemerythrin and azidomethemerythrin at 2·0 Å resolution. *Acta Crystallogr. sect. B*, **39**, 697–703.

Stout, C. D. (1989). Refinement of the 7 Fe ferredoxin from *Azotobacter vinelandii* at 1·9 Å resolution. *J. Mol. Biol.* **205**, 545–555.

Tainer, J. A. Getzoff, E. D., Beem, K. M., Richardson, J. & Richardson, S. D. C. (1982). Determination and analysis of the 2Å structure of copper, zinc superoxide dismutase. *J. Mol. Biol.* **160**, 181–217.

Takano, T. (1984). Refinement of myoglobin and cytochrome *c*. In *Methods and Applications. Crystallographic Computing* (Hall, S. R. & Ashida, T., eds), pp. 262–272, Oxford University Press. Oxford, England.

Terwilliger, T. C. & Eisenberg, D. (1982). The structure of melittin. I. Structure determination and partial refinement. *J. Biol. Chem.* **257**, 6010–6015.

Tsernoglou, D. & Petsko, G.A. (1977). Three-dimensional structure of neurotoxin A from venom of the Philippines sea snake. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 971–974.

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.

Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, **6**, 1425–1436.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). Structure of ubiquitin refined at 1·8 Å resolution. *J. Mol. Biol.* **194**, 531–544.

Weaver, L. H. & Matthews, B. W. (1987). Structure of bacteriophage T4 lysozyme refined at 1·7 Å resolution. *J. Mol. Biol.* **193**, 189–199.

Wistow, G., Turnell, B., Summers, L., Slingsby, C., Moss, D., Miller, L., Lindley, P. & Blundell, T. (1983). X-ray analysis of the eye lens protein gamma-II crystallin at 1·9 Å resolution. *J. Mol. Biol.* **170**, 175–202.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, 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.

Wlodawer, A., Savage, H. & Dodson, G. (1989). Structure of insulin. Results of joint neutron and X-ray refinement. *Acta Crystallogr. sect B*, **45**, 99–107.

Wright, C. S. (1987). Refinement of the crystal structure of wheat germ agglutinin isolectin 2 at 1·8 Å resolution. *J. Mol. Biol.* **194**, 501–529.

Zhang, R. G., Joachimiak, A., Lawson, C. L., Schevitz, R. W., Otwinowski, Z. & Sigler, P. B. (1987). The crystal structure of Trp apo-repressor at 1·8 Å shows how binding tryptophan enhances DNA affinity. *Nature (London)*, **327**, 591–597.

Edited by A. Klug