Acta Cryst. (1994). D**50**, 889–899

Comparison of Homology Models with the Experimental Structure of a Novel Serine Protease

BY MIKE CARSON

Center for Macromolecular Crystallography, University of Alabama at Birmingham, Birmingham, AL 35294, USA

CHARLES E. BUGG

BioCryst Pharmaceuticals, Inc., 2190 Parkway Lake Drive, Birmingham, AL 35244, USA

AND LAWRENCE J. DELUCAS AND STHANAM V. L. NARAYANA

Center for Macromolecular Crystallography, University of Alabama at Birmingham, Birmingham, AL 35294, USA

(*Received* 13 *October* 1993; *accepted* 5 *May* 1994)

Abstract

A model structure of the human complement enzyme factor D was built based on homology with related serine proteases. A molecular-replacement solution of the factor D crystal structure employing the homology model refined without manual intervention to an *R* factor of 0.249 with 2.4 Å native diffraction data. A multiple isomorphous replacement (MIR) electron-density map was subsequently produced, leading to a model refined at 2.0 Å resolution to an *R* factor of 0.188. A homology model built with commercial modeling software was subjected to the same procedure. Comparisons of the homology models with the final refined MIR structure are presented. Major discrepancies were found in critical active-site regions.

1. Introduction

There is a great deal of interest in creating models of proteins based on sequence homology with known structures. These models may be used on their own, or may provide starting points for refinement against experimental data. A prerequisite of this approach is the availability of accurate, experimentally determined atomic structures of at least one, and preferably many, related proteins. Some examples of this approach include the modeling of renin (Sibanda *et al.*, 1984), immunoglobulins (Chothia *et al.*, 1989) and serine proteases (Greer, 1990).

Factor D is the first serine protease from the complement system to be crystallized (Narayana *et al.*, 1991) and to have its structure determined at atomic resolution (Narayana *et al.*, 1994). It is a single polypeptide chain of 228 residues. Factor D is the rate-limiting enzyme in the 'cascade' of events that activate and regulate the alternate pathway of the complement system (Lesavre & Muller-Eberhard, 1978).

We describe the creation of an '*ad hoc*' model of factor D by computer graphics modeling and the 'solution' of the structure by molecular replacement using this homology model. The subsequent creation and refinement of a second homology model built using a commercial software package is also described.

The accuracy of models constructed from homology is of fundamental concern when the models cannot be confirmed by experimental methods. In addition, there is some question about the general validity of refined crystal structures obtained through molecular-replacement techniques employing homology models. In order to understand better the errors that might arise from the use of homology models, we have investigated the differences between the structure of human complement factor D determined from MIR crystallographic refinement and the homology models built from known structures of related serine proteases.

2. Methods

2.1. *Homology modeling*

Homology modeling was used to create a full-atom model of factor D based on previously determined proteins structures. The Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977) provided the data.

2.2. *Ad hoc modeling*

This methodology was inspired by a talk given by Greer (1988).

2.2.1. *Sequence homology.* Sequence homology searches were performed with the program package *IDEAS* (integrated database and extended analysis system for nucleic acids and proteins) running on the Alabama supercomputer network Cray X-MP employing the *FASTA* algorithm (Pearson & Lipman, 1988). A custom database was constructed with all X-ray structures in the January 1990 release of the PDB. The default parameters recommended for alignment were used.

2.2.2. *Model building.* Structures were superimposed manually employing the program *PSFRODO* (Pflugrath, Saper & Quiocho, 1984). Serine proteases are predominantly β-sheet structures, with two small α-helices and four conserved disulfide bonds. One C_α tracing served as

the reference; others were visually fitted in stereo to best align the highly conserved secondary structural features.

The model was created from a related sequence that was changed, with appropriate insertions and deletions, using options in *FRODO* (Jones, 1978). Missing atoms were added with the *REFINE* option (Hermans & McQueen, 1974) and a dictionary modified to use the side-chain rotamer conformation occurring most frequently in highly refined protein structures (Ponder & Richards, 1987). This crude model was modified on a Silicon Graphics IRIS workstation using the program *Atom*. *Atom* (Alabama *TOM*) is a local variant of *TOM* (Cambillau & Horjales, 1987), itself a variant of *FRODO*. A customized user interface with pop-up menus allows selection of most probable side-chain and main-chain conformations. Pointing and clicking on a φ/ψ plot invokes refinement to a particular main-chain conformation.

Side-chain conformations were not modified in the case of an exact match with the template structure (about 35% of the residues). The remaining side-chain conformations were selected to best mimic the conformation in one of the overlaid structures. For example, if a Phe of the template became a Leu in factor D, the Leu rotamer which best overlapped its C_δ atoms with those in the Phe ring was selected. In ambiguous cases, selections were based on well known principles of protein structure: form hydrogen bonds if possible, place polar groups outside and hydrophobic groups inside, and best fill space.

Geometric regularization was carried out with the *REFINE* options of *FRODO*. A model was checked by calculating its potential energy with *X-PLOR* (Brünger, Kuriyan & Karplus, 1987).

2.3. *Automated modeling*

A commercial molecular modeling package became available for trial at a later date. A model of factor D was created to test this system.

2.3.1. *Software features*. The *Protein Design* module of *Quanta* (Polygen Molecular Simulations, Incorporated, 200 Fifth Avenue, Waltham, MA 02254) is designed to create homology models as briefly de-

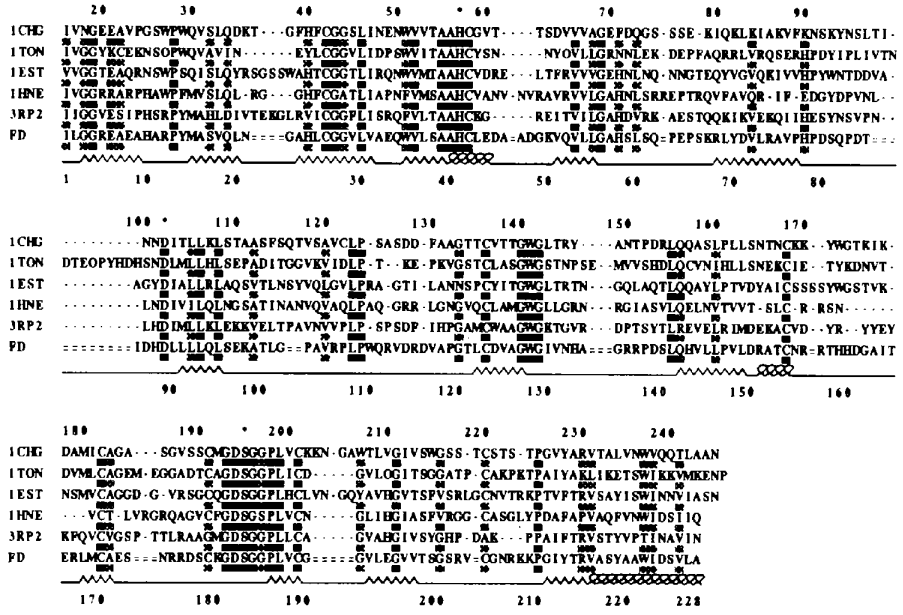

Fig. 1. Alignment of factor D with selected serine proteases. The primary structures used for modeling (Tables 2 and 3) are listed by their PDB codes. Dashes mark where one sequence has insertions relative to the other. Black (exact) and gray (nearly exact) mark homologous residues. The 'chymotrypsinogen' numbering scheme is given in the top row of numbers. Asterisks mark the active-site triad. The factor D sequence is shown at the bottom, and is numbered consecutively from 1 to 228. The secondary structure of factor D is shown schematically with loops for helices, zigzags for sheets and lines for coils or turns.

Fig. 2 Superposition of serine proteases. *B*-spline ribbon drawings (Carson & Bugg, 1986) for the seven serine proteases of Table 1. A thick line denotes sheet or helix, thin lines represent turns or coils. Residue labels in the chymotrypsinogen scheme are shown for the termini and active-site triad. Chymotrypsinogen, a pro-enzyme, has 15 additional residues at the N-terminus.

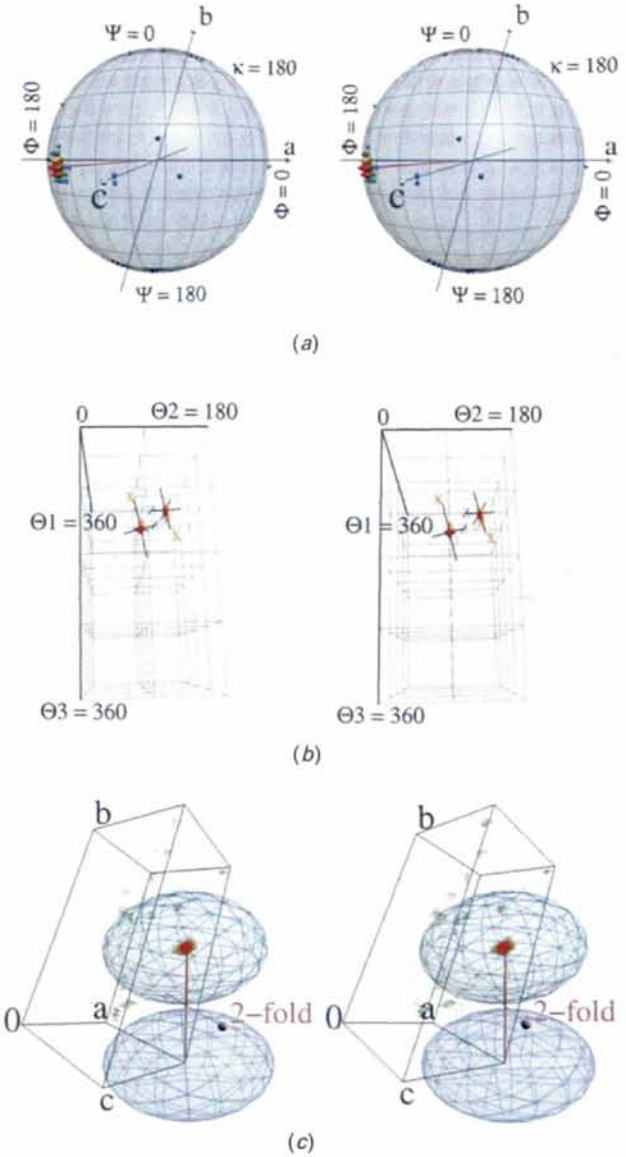

Fig. 3. Molecular-replacement solution of factor D. (*a*) Self-rotation function. The unit-cell axis vectors (**a**, **b**, **c**) are superimposed on the unit sphere. The polar-angle values φ (longitude) and ψ (latitude) are shown in 15° increments. All peaks in the rotation function of at least 2 standard deviations (σ) above background are shown as cones. The size and shading of the cones reflect the value of the function. A vector from the origin to the largest peak marks the direction of the observed pseudo-twofold. (*b*) Cross-rotation function. The unique Eulerian angle space is represented as a three-dimensional grid. All peaks of at least 4σ are shown as semi-transparent cubes. Axes labeled *X*, *Y*, *Z* are centered at the two observed peaks. The axes have been rotated by the Eulerian angles of the peak. It may be seen that these peaks are related by a twofold. (*c*) Translation function. The unit-cell box is shown with the origin and axes labeled. All peaks in the *R*-factor search translation function of at least 3σ are shown as semi-transparent cubes. The size and shading of the cubes reflect the value of the function. Each monomer is represented as an ellipsoid fit to the model structure. A vector from the largest peak to the closest origin (+*X*, +*Z*) marks the solution. The twofold axis determined from the self-rotation function is also displayed at the midpoint of the translation vector. This point was chosen as the origin of the cell (arbitrary in *P*1) to simplify the non-crystallographic symmetry analysis.

scribed below. Secondary structural assignments are made based on the structure of known proteins (Kabsch & Sander, 1983), sequence alignments are made with the *FASTA* program (Pearson & Lipman, 1988), searches of files in PDB format are allowed, and superposition of coordinates may be carried out in a simple fashion by combining sequence match and secondary structure match information. The homology model is built by copying conformations and sequences between proteins, while performing insertions, deletions and mutations of residues as needed. Automatic change of side-chain conformations is carried out to remove or minimize bad contacts. Energy minimization with constraints using *CHARMm* (Brooks *et al.*, 1983) completes the process.

The *Protein Design User's Guide*, a tutorial, describes the steps required to model the human renin protein. The renin sequence is known, and has been modeled based on available structures of homologous proteins (Sibanda *et al.*, 1984). This tutorial was followed to create a model of human factor D in an automated fashion using all the recommended default values and procedures.

2.4. *Crystallographic analysis*

X-PLOR Version 2.1 was employed for all aspects of the initial crystal structure solution, refinement, and analysis. *X-PLOR* Version 3.0 was used in the later stages.

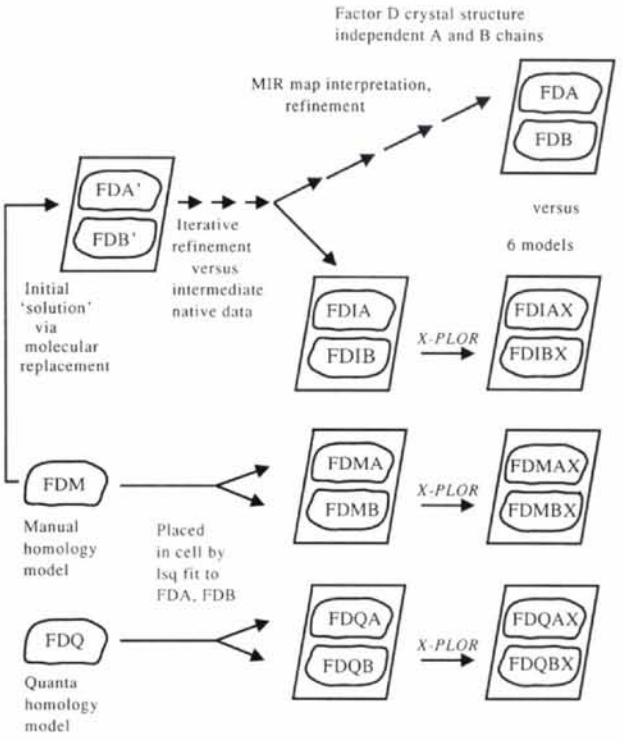

Fig. 4. Flowchart of model creation and nomenclature. *X-PLOR* refers to one round of SA refinement with *X-PLOR* against the final native data. No manual adjustments were made on any of the six refined models.

2.5. *Pairwise comparison*

The two factor D models to be compared are first superimposed with *X-PLOR* by a least-squares fitting of all the C_α atoms. An *X-PLOR* script computes root-mean-square (r.m.s.) differences in atomic coordinates (Å) between two models on a per residue basis, taking into account the symmetry of Asp, Glu, Tyr and Phe residues. The difference between two structures may also be expressed in terms of dihedral angles. The differences in φ/ψ or $\chi 1/\chi 2$ pairs are computed as Euclidian distances in radians, *i.e.* for the main chain, $[(\Delta\varphi)^2 + (\Delta\psi)^2]^{1/2}$. This is called the dihedral difference.

Plots are made on a per-residue basis, considering main-chain and side-chain atoms separately. The C_α atom is included in both the main-chain and side-chain r.m.s. computations. Residue numbering is based on consecutive integers from 1 to 228. (See Fig. 1 for comparison with the chymotrypsinogen convention.)

The structure of sarcoplasmic calcium-binding protein (SCP) (Vijay-Kumar & Cook, 1992) was analyzed for comparison. SCP is a helical protein solved in this laboratory by similar methods at the same resolution to a similar *R* factor. Two monomers of 174 residues are in the asymmetric unit. The error suggested by Luzzati (1952) plots is 0.23 Å. In this case, 40% of the main-chain residues after superposition were within 0.23 Å and 82% were within 0.46 Å. The values were 20 and 55% for the side chains. A total of eight main-chain and 49 side-chain residues differ by more than 1.0 Å.

A standard to monitor the agreement between two structures is required for the analysis that follows. We adopt the criterion that a model agrees if the r.m.s. deviation is within twice the error suggested by the Luzzati plots.

3. Results

3.1. *Ad hoc model building*

The results of the best homology scores, the exact sequence match percentage, and a description of the protein are given in Table 1. The top seven unique structures (all members of the serine protease family) are shown sorted by the calculated homology score. Each structure is referenced by its PDB identifier: 3RP2 (Remington, Woodbury, Reynolds, Matthews & Neurath, 1988), 1HNE (Navia *et al.*, 1989), 1EST (Sawyer *et al.*, 1978), 1TON (Fujinaga & James, 1987), 1CHG (Freer, Kraut, Robertus, Wright & Xuong, 1970), 2TRM (Sprang *et al.*, 1987), 1TGB (Fehlhammer *et al.*, 1977). The results of the alignment are shown in Fig. 1.

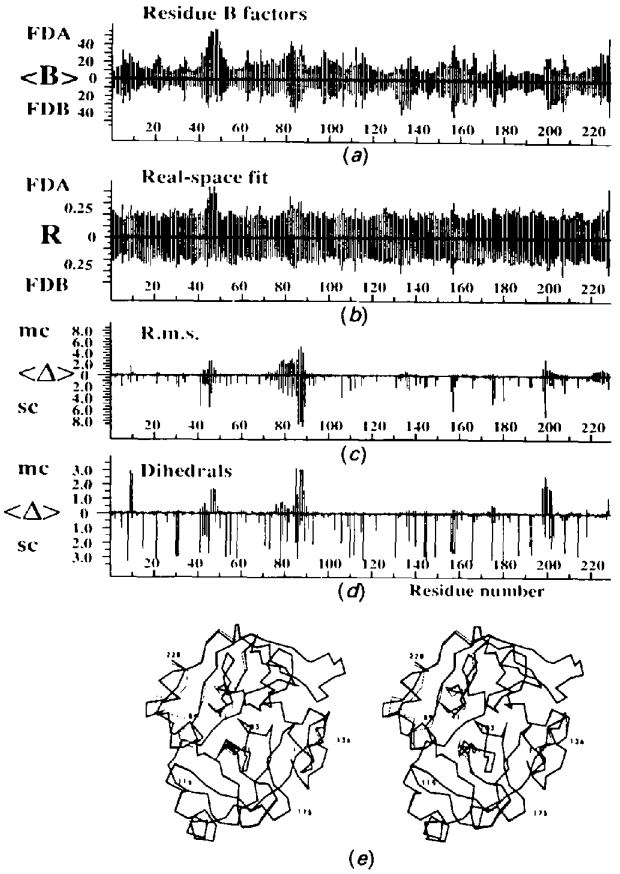

Fig. 5. Differences between FDA and FDB. (*a*) The average *B* factor of each residue. (*b*) The real-space fit residual of each residue. (*c*) The r.m.s. difference (Å) for the main-chain (mc) and side-chain (sc) atoms for each residue of the superimposed structures. (*d*) The dihedral-angle differences in radians. (*e*) Stereo C_α tracings of the superimposed structures. FDB is shown as the thicker lines.

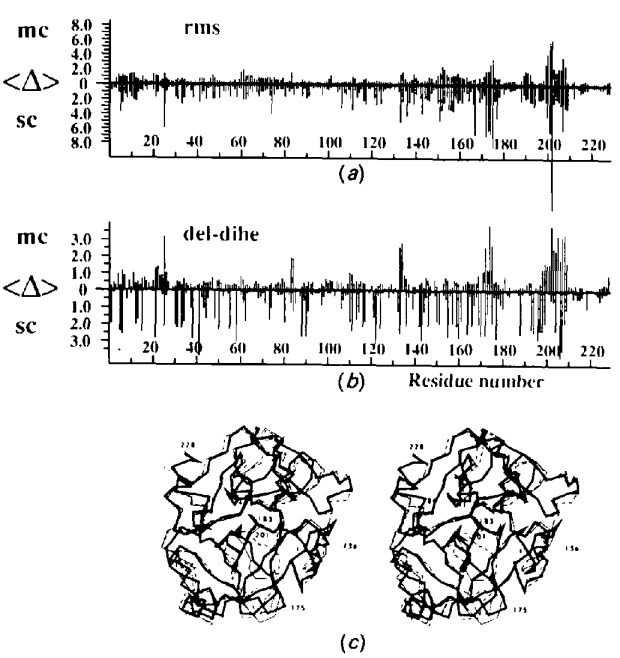

Fig. 6. Differences between FDM and FDQ. (*a*) The r.m.s. difference (Å) for the main-chain (mc) and side-chain (sc) atoms for each residue of the superimposed structures. (*b*) The dihedral-angle differences in radians. (*c*) Stereo C_α tracings of the superimposed structures and the crystal structure FDB. FDB is shown as the thicker lines, FDM as thin lines and FDQ as dashed lines.

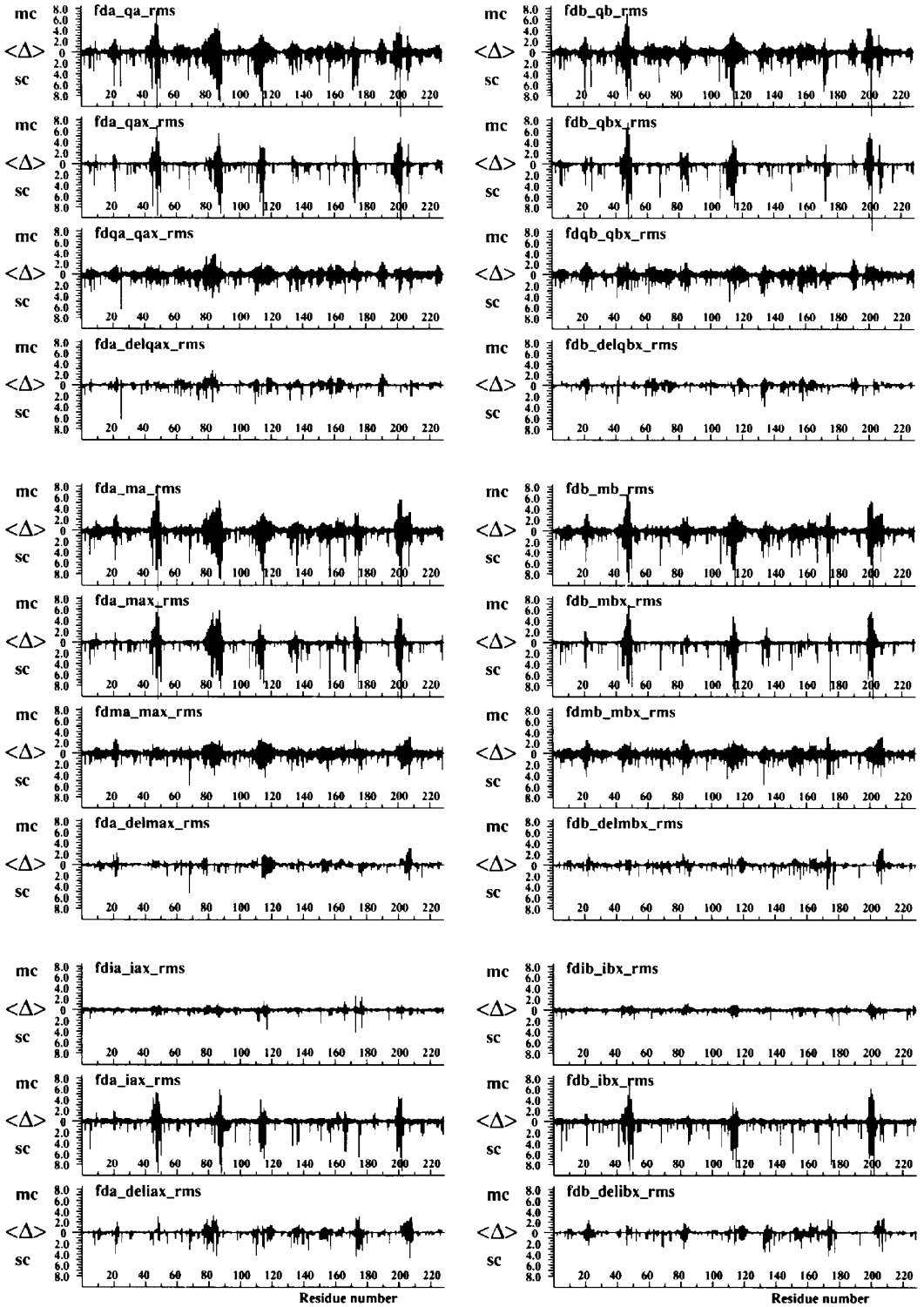

Fig. 7. Differences between crystal structure and models. Deviations and shifts of the models (Å) for for the main-chain (mc) and side-chain (sc) atoms on a per-residue basis. The results for FDA comparisons are on the left and on the right for FDB. The nomenclature is explained using FDM as an example. fda_ma is the difference between the crystal structure and the initial homology model. fda_max is the difference between the crystal structure and the homology model after SA refinement. fdma_max is the shift resulting from the refinement. The improvement resulting from *X-PLOR* refinement is fda_delmax, defined as fda_max minus fda_ma. The few negative values are set to zero (see Table 6). The FDI models had FDM as their starting point. For these, only deviations from the previous iteration (*e.g.* fdia_iax) are shown.

Table 1. *Sequence homology with factor D*

The proteins are sorted by their *IDEAS* homology score. References to the PDB files are given in the text.

PDB code	Homology score	% Exact match	Protein description
3RP2	305	33.9	Rat mast cell protease
1HNE	283	33.5	Human neutrophil elastase
1EST	282	33.5	Porcine pancreas elastase
1TON	238	34.4	Rat submaxilliary gland tonin
1CHG	178	33.3	Bovine chymotrypsinogen-A
2TRM	173	33.8	Rat trypsin mutant
1TGB	164	35.5	Bovine pancreas trypsinogen

Table 2. *Alternative factor D model templates*

The 3RP2 structure was employed for the remainder of the 228-residue sequence in constructing the manual model.

Range	Protein
20–26	1HNE
43–50	1HNE
155–170	1EST
171–178	1CHG
198–208	1HNE

The sequences show a high degree of homology. The three-dimensional similarity is even more striking. The C_α tracing of 3RP2 served as the reference. The resulting superpositions are shown in Fig. 2. The major differences in backbone structure occur in the loops connecting secondary structural elements. These loop regions are also where the 'insertions' and 'deletions' in the alignment occur.

The factor D model was created using 3RP2 as the basic template. Each of the 228 residues were examined sequentially from the N to the C terminus. The atomic models of 3RP2 and the two or three next best sequence matches for each stretch of approximately 20 residues were displayed. There was always at least one other protein with the same number of residues as factor D in regions where insertions/deletions were made in 3RP2. The regions of factor D where 3RP2 was not the primary template are given in Table 2. Approximately 50 residues were shifted to align the main chain (including carbonyl groups) with the alternate protein. All main-chain conformations were set to allowed φ/ψ values. All side-chain conformations were set to favored rotamers.

The potential energy of the model was calculated with *X-PLOR*. The geometry was nearly ideal. Two pairs of residues elicited warnings of atoms being less than 1.5 Å apart. These positions were re-examined with graphics and deemed inconsequential: one formed a hydrogen bond and the other a non-polar interaction. Energy refinement by 250 cycles of conjugate gradient minimization with *X-PLOR* was performed. The r.m.s. changes from the model built with *Atom* to the energy minimized model were 0.60 Å for the 228 C_α atoms and 0.91 Å for all 2113 atoms. This produced the model fdm_6jun90.pdb, which is referred to as FDM (factor D from manual modeling).

3.2. *Automated model building*

The PDB was not loaded on this trial system. The first step involving a complete search of the PDB was omitted as the method also uses *FASTA* for alignment. The seven previously identified proteins of Table 1 were imported into this modeling package's environment.

A multiple sequence alignment (Feng & Doolittle, 1987) aligned all sequences to that of factor D. The resulting alignments were excellent, correlating well with secondary structure. Examination of the alignment scores and tutorial recommendations indicated that factor D should be built from 1TON, 3RP2, 1HNE and 1EST. A two-residue deletion exists in factor D between residues 174 and 175; otherwise, there was always an alignment match. Table 3 gives the residue ranges from those proteins used to create the homology model.

The coordinates of the known structures were copied onto the sequence of factor D. Exact sequence matches were copied directly; otherwise, only atoms in common between the two residue types were used. The regularization routine added the missing atoms. Residues that were joined together from different protein frag-

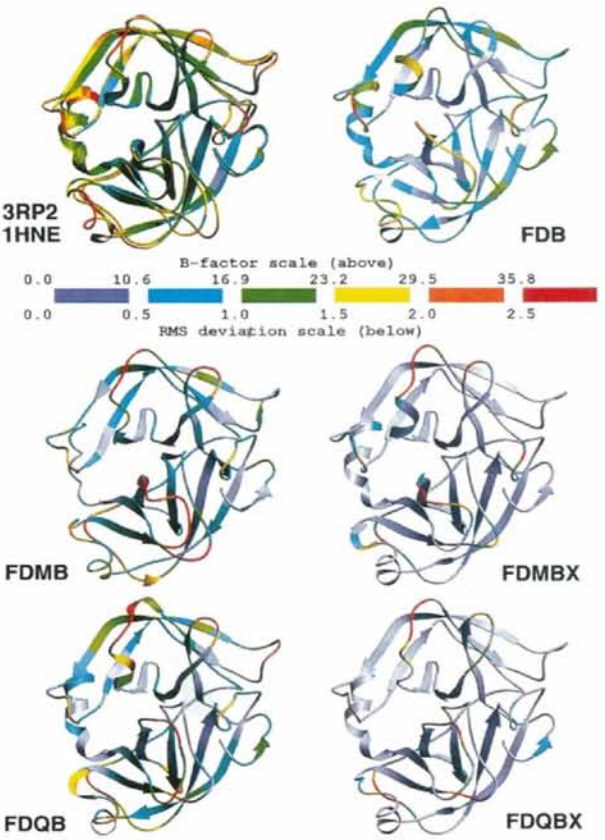

Fig. 8. Factor D model *B* factors and r.m.s. deviations. The two serine proteases, 3RP2 and 1HNE, most used in the homology modeling and the final FDB chain are color coded per residue by r.m.s. main-chain *B* factor. The r.m.s. deviations from the final FDB structure are also given. The initial homology models of factor D are FDMB and FDQB. These models after the *X-PLOR* SA protocol are FDMBX and FDQBX.

Table 3. *Protein design factor D model templates*

The automated model was built piecewise from the fragments shown. Note that factor D has a two-residue deletion at position 174 compared to the 1TON sequence.

Range	Protein
1–24	1HNE
25–42	1TON
43–77	1HNE
78–132	3RP2
133–167	1EST
168–174	1TON
175–209	1TON
210–228	3RP2

ments (which may introduce gaps) were examined and deemed to have reasonable geometries. The two-residue deletion mentioned above was in a turn extending into the solvent far from the active site. The regularization routine annealed these gaps. The side-chain spin routine checked all side-chain conformations for bad contacts. Energy minimization of the entire structure produced the model fdq_24oct91.pdb. This model is referred to as FDQ (factor D from *Quanta*).

3.3. *Crystallographic results*

3.3.1. *Molecular-replacement solution.* Factor D crystallizes in the triclinic space group P_1 with two independent molecules per asymmetric unit. The original homology model, FDM, was used to solve the crystal structure of factor D by molecular replacement (Rossmann, 1972). The various functions gave unique peaks considerably above background (Fig. 3).

3.3.2. *Initial structure refinement.* Rigid-body refinement of the two independent monomers produced little change. Examination of graphics revealed no interpenetration as a result of crystal packing, although some residues were too close. Refinement continued with no manual adjustment of the coordinates or non-crystallographic symmetry restraints. All refinements included reflections greater than 2σ from 7.5 Å to the high-resolution limit and an overall temperature factor of 15 Å^2. The recommended *X-PLOR* slow-cooling simulated-annealing (SA) protocol was followed. The *R* factor decreased from 0.455 to 0.218 at 3.0 Å resolution (5657 reflections) producing the initial molecular-replacement model, fdx_29jun90.

Additional rounds of SA refinement were undertaken as improved native data sets became available. The 2.5 Å data set (11 220 reflections) gave a model (fdx_24oct90) with an *R* factor of 0.242. These coordinates established the origin and proved very useful in locating heavy-atom derivative positions. The 2.4 Å data (13 032 reflections) yielded a model (fdx_19jan91) with an *R* factor of 0.249. This was the starting point for interpreting the first MIR maps.

3.3.3. *Initial error analysis.* Comparison of the intermediate fdx_24oct90 and the fdx_19jan91 coordinates indicated that 36 main-chain residues and 39 side-chain residues of the 456 independent residues had moved significantly (> 1 Å for main-chain residues, > 2 Å for side-chain residues). A $2F_{obs} - F_{calc}$ map was created with phases calculated from the latter model. The real-space fit residual (Jones, Zou, Cowan & Kjeldgaard, 1991) was computed for the main chain and side chain of each residue. Geometric analysis revealed residues with unfavorable dihedral angles.

3.3.4. *Initial MIR map.* The initial MIR map was computed at 3.2 Å using three heavy-atom derivatives ($PtCl_4$, PCMBS, K_2HgI_4), and symmetry averaged (SVLN, 30jan91). Preliminary inspection indicated that approximately 80% of the model coordinates fit the observed density very well. An additional 10% of the residues required only minor adjustment. The remaining 10% of the residues had problems; about half required major adjustment (a few ångströms shift) and the rest were uninterpretable in this map. These 'problem' residues corresponded well to those identified by the initial error analysis. These preliminary results were published in conference proceedings (Carson *et al.*, 1991).

3.3.5. *Crystallographic refinement.* The details of the crystallographic refinement of factor D to 2.0 Å are given by Narayana *et al.* (1994). We stress that phases from the homology model were never used (except in aiding the location of heavy-atom positions). We believe the refined model is of high quality. These coordinates are thus assumed to be correct and, therefore, provide the basis for the comparisons that follow.

3.4. *Model nomenclature*

The final refined crystal structure models of the two independent factor D monomers are referred to as 'FDA' and 'FDB'. The initial homology models are referred to as 'FDM' and 'FDQ' for the manually (M) constructed and the *Quanta*-generated (Q) coordinates, respectively.

The molecular-replacement model (fdx_19jan91) was subjected to SA and temperature-factor refinement against the final 2.0 Å data set of 23 249 reflections. Thus, this model has now undergone five iterations (I) of *X-PLOR* refinement after its initial placement in the unit cell. The individual subunits are denoted as FDIAX and FDIBX.

The original FDM and FDQ models were placed in the unit cell by a least-squares fit with the corresponding C_α atoms of FDA and FDB. These models were then subjected the SA protocol of *X-PLOR* (X) described previously. Additionally, individual atomic temperature factors were refined. This produced the models FDMAX, FDMBX, FDQAX and FDQBX for the two monomers refined from each starting homology model. (This assumes that the molecular-replacement solutions would be found exactly. The ease of solution with the initial FDM model using incomplete data makes this plausible.) Even though both the FDMX and FDIX models used

Table 4. *R factors of final and molecular replacement models*

All molecular replacement models have undergone *X-PLOR* SA refinement as described in the text. *R* factors are based on 7.5–2.0 Å native data with a 2σ cutoff. The bonds (Å) and angles (°) columns are the r.m.s. deviations from ideal geometric values. The errors (Å) are estimated from Luzzati plots (not shown).

Model	*R* factor	Bonds	Angles	Error	Description
FD	0.188	0.010	1.65	0.23	Final crystal structure
FD'	0.219	0.010	1.65	0.27	Final, 69 waters omitted
FDI	0.246	0.017	2.23	0.30	Iterated manual model
FDQ	0.255	0.017	2.29	0.32	*Quanta* model
FDM	0.259	0.018	2.28	0.32	Manual model

FDM as the starting point, the r.m.s. differences between the closest pair, FDIBX/FDMBX, is 0.8 Å for main-chain atoms and 1.7 Å for all atoms. Thus, they are considered distinct models. None of the models refined by *X-PLOR* had any manual intervention. A flow chart of the models' creation and nomenclature is given in Fig. 4.

3.5. *R factors and estimated error*

The final *R* factors, deviations from ideal values, and estimated coordinate errors for the models are given in Table 4. The models are sorted by *R* factor. The experimental crystal structure clearly provides the best model, but the differences between it and the homology models are not dramatic. The estimated errors are all less than 0.33 Å.

3.6. *Comparison of the two independent crystallographic monomers*

The two subunits in the crystal were refined independently, with some interesting conformational differences noted between FDA and FDB (Narayana *et al.*, 1994). Analysis of the crystal structure reveals that the FDB monomer more closely fits the experimental data, in particular the range of residues from 41 to 48. These residues are disordered in the final FDA structure. The refined temperature factors and real-space fit per residue are shown in Fig. 5.

The main-chain dihedral differences of 202 of the 228 residues agree within 30°. Only 16 residues have differences over 60°. There are 159 side-chain residues with dihedral differences of less than 30°, and 57 residues with differences greater than 60° (implying a different rotamer). The dihedral differences between FDA and FDB and the r.m.s. differences between the superimposed monomers are plotted in Fig. 5. A C_α tracing of the superposition is also shown.

Slightly more than half of the 228 residues show main-chain deviations less than the 0.23 Å error suggested by Luzzati plot analysis. Slightly more than three quarters have deviations no greater that twice that amount (0.46 Å). The corresponding results for side-chain residues are 29 and 53%, respectively. There are 27 main-chain and 76 side-chain residues with r.m.s. differences over 1.0 Å. From visual inspection, there are significant main-chain conformational differences in the ranges 41–48, 77–89 and 198–201. 20 of the side chains that differ significantly have similar main-chain conformations. These are primarily surface Arg, Lys and Glu influenced by crystal packing.

3.7. *Comparison of the two different homology models*

The FDM and FDQ models were created independently, but are based on the same set of protein templates. However, comparison of Tables 2 and 3 indicates that less than half of the residues (43–50 1HNE, 78–132 3RP2, 155–167 1EST, 210–228 3RP2) had an identical protein as the primary template.

Fig. 6 plots the dihedral differences between FDM and FDQ and the r.m.s. differences between FDM and the superimposed FDQ. A C_α tracing of the superpositions on the crystal structure FDB is also shown. There are only 71 main-chain and 28 side-chain residues with r.m.s. deviations less than the 0.46 Å cutoff. There are 61 main-chain and 125 side-chain residues with deviations over 1.0 Å. The overall average r.m.s. difference is 1.2 Å for the main-chain atoms and 2.6 Å for all atoms. Comparison of dihedral angle differences using a 30° cutoff shows similar conformations for 145 main-chain and 135 side-chain residues. A 60° cutoff yields significant differences for 38 main-chain and 77 side-chain conformations.

3.8. *Comparison of the crystal structure with models*

Comparisons with the appropriate FDA or FDB are presented in the upper half of Table 5 and in Fig. 7. There is little difference between the degree of fit of the FDM and FDQ models to the crystal structure. Both fit slightly better to FDB than to FDA. Less than 20% of the main chain and less than 10% of the side chains may be considered correct based on the difference in coordinates. However, examination of the differences in dihedral angles indicates that approximately 60% of the main-chain and 50% of the side-chain conformations are nearly correct. The differences are not uniformly distributed; about half of the large differences occur in the regions where FDA differs from FDB.

3.9. *Comparison of crystal structure with refined models*

Comparison of the homology models after SA refinement is given in the lower portion of Table 5 and in Fig. 7. The *X-PLOR* refinement moved the coordinates significantly closer to the crystal structure for almost every residue. These results are summarized in Table 6.

Examination of Fig. 7 reveals FDI is marginally better than the other models. This model also shows the smallest changes caused by *X-PLOR*, as it was subjected to four prior rounds of SA refinement.

3.10. *Comparison summary*

Both model structures fail to reproduce accurately the crystal structure, especially around the active site

Table 5. *Deviations of models from crystal structure*

The main-chain (mc) and side-chain (sc) r.m.s. deviations over the entire structure (Δ), and the median value (med) of the deviations over all 228 residues are given in Å. The mc-ok and sc-ok count residues having deviations within twice the Luzzati limit. The $\varphi\psi$-ok and χs-ok count dihedral differences of less than 30°

Model	mc-Δ	sc-Δ	mc-med	sc-med	mc-ok	sc-ok	$\varphi\psi$-ok	χs-ok
FDMA	1.73	3.55	0.75	1.51	37	15	131	114
FDMB	1.56	3.22	0.74	1.45	41	14	136	116
FDQA	1.60	3.40	0.80	1.76	52	25	142	106
FDQB	1.44	3.02	0.78	1.57	54	21	153	110
FDMAX	1.44	3.12	0.30	0.69	165	85	166	126
FDMBX	1.17	2.81	0.28	0.53	184	111	187	132
FDQAX	1.27	2.88	0.26	0.65	174	96	177	130
FDQBX	1.27	2.84	0.26	0.66	177	96	175	130
FDIAX	1.22	2.70	0.44	0.59	124	74	179	136
FDIBX	1.14	2.68	0.43	0.62	139	68	189	132

Table 6. *X-PLOR shifts of models*

The number 'better' counts residues that moved closer to the refined crystal structure after *X-PLOR* SA refinement. The maximum shifts for any residue are given in Å. The number 'worse' counts residues that moved more than twice the Luzzati limit away from the crystal structure. The number 'bad' counts residues with r.m.s. deviations greater than 1.0 Å.

Model protein	No. mc better	No. sc better	max-mc Å shift	max-sc Å shift	No. mc worse	No. sc worse	No. mc bad	No. sc bad
EDMAX	201	186	3.06	5.41	9	20	42	101
FDMBX	208	187	3.09	4.59	2	17	36	98
FDQAX	210	191	2.83	6.33	4	13	33	85
FDQBX	201	183	1.91	4.01	10	28	25	86
FDIAX	187	184	3.32	4.76	4	8	34	96
FDIBX	184	171	2.78	4.43	1	2	23	85

and substrate-binding loops. Only about 10% of the molecular-replacement model main chain was grossly in error after the SA refinement. Almost all of these errors were located in the active-site and substrate-binding regions of the enzyme. These are precisely the residues that must be known accurately to understand the structure/function relationship for this enzyme. However, these loops are likely to be rather flexible. The temperature factors of the refined crystal structures of factor D and the serine proteases used for homology modeling are generally higher in these loops. The results are summarized in Fig. 8.

4. Discussion

4.1. *Confidence in models*

A model of factor D was built using interactive graphics, based on visual homology modeling with seven superimposed serine proteases. The model had nearly ideal geometry, all dihedral angles set to allowed values, and violated none of the 'rules' of protein structure. The modeler's naive view, perhaps seduced by the beauty of computer graphics, was that there might be no need to perform the experiment.

An unambiguous molecular-replacement solution and subsequent refinement against the native data produced very respectable molecular geometry and *R* factor, at the expense of substantial coordinate shifts. The modeler's revised naive view held that the structure was essentially solved, and refinement could be completed with a few rounds of refitting based on difference maps using native data only.

The view of the experimentalists holds that one might not escape from the bias of the model. Experimental MIR phases were determined; phases from the molecular-replacement model were never used. A high-quality 2.0 Å structure was ultimately produced. The differences between the final structure and the homology models have been documented here. These differences often exceed what most crystallographers would consider correct.

4.2. *Modeling methods*

The production of the manual model (FDM) required two days work on a high-performance graphics workstation. The production of the more automated model (FDQ) with commercial software required almost a day. (Another protein of similar size required only 2 h to model, as familiarity with the software was attained.)

FDM and FDQ are roughly equidistant from the true structure, both before and after refinement. They are nearly as different from each other as they are from the crystal structure. Both model structures fail to accurately reproduce the crystal structure around the active-site substrate-binding loops. An experienced researcher can build as good a model with freely available academic software as with a commercial package. However, the commercial package has a better user interface, integrates more features, and can accomplish the task more quickly.

Few systems should be as easy to model as serine proteases, given the wealth of data on these proteins. The models built purely by graphical methods have severe errors. The final crystal structures were examined after the fact to determine if the regions of poor fit to the homology models could have been modeled better to begin with. A variant of the 'spare parts' method (Jones & Thirup, 1986) was used with a database of 62 highly refined protein structures. Residues 199–202 should have been modeled with coordinates from 3EST. Several proteins fit the bend from 113–116 better, but none of the proteases. The longer stretches 43–50, 81–89, and 161–167 generally had some protein that would fit to an r.m.s. of 1–1.5 Å, but it is unclear how these might have been selected in the first place.

Irrespective of the source of the model, experimental data will be required to back it up. The improvement in the models after *X-PLOR* refinement against the empirical data is impressive (consult Tables 4 and 5 and Figs. 7 and 8). However, the structure apparently becomes trapped in a local minima from which it cannot escape, even after many iterations of annealing.

4.3. *Future work*

We do not wish to cast doubt on the usefulness of molecular replacement in general. However, there is a major unanswered question: could the correct structure have been attained without resorting to MIR methods and many rounds of manual refitting of graphics?

Fig. 9 shows the completely computer-generated FDIBX molecular-replacement model with the FDB crystal structure and the computed OMITMAPS based only on the FDIBX model and the native data. It would appear that the model could be refitted to this map, *especially if one knew that this particular region was in error.*

Fig. 9. True *versus* model structure and OMITMAP. The final crystal structure FDB (bold, left) and the refined model structure FDIBX (medium, right) are shown as line drawings between residue 45 and 49, the first point where there is a serious deviation between the two, and the map no longer fits the model well. The OMITMAP (Bhat & Cohen, 1984), contoured at 0.67σ units of the map, was based entirely on the 2.0 Å native data and the FDIBX model.

We have developed a statistical protocol that can largely identify the incorrect residues with few false positives in an accompanying paper (Carson, Buckner, Yang, Narayana & Bugg, 1994). The protocol employs temperature factors, real-space fit residual, geometric strain, dihedral-angle sensibility and coordinate shifts from the previous refinement cycle. We intend to have a student unaware of the history of this project attempt to refit the problem residues with map-fitting software under development.

We gratefully acknowledge NASA grant NAGW-813 and Public Health Sevice grant AI32949 for support.

References

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). *J. Mol. Biol.* **112**, 535–542.

Bhat, T. N. & Cohen, G. H. (1984). *J. Appl. Cryst.* **17**, 244–248.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comp. Chem.* **4**, 187–217.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science* **235**, 458–460.

Cambillau, C. & Horjales, E. (1987). *J. Mol. Graphics*, **5**, 174–177.

Carson, M., Buckner, T. W., Yang, Z., Narayana, S. V. L. & Bugg, C. E. (1994). *Acta Cryst.* D**50**, 899-908.

Carson, M. & Bugg, C. E. (1986). *J. Mol. Graphics*, **4**, 121–122.

Carson, M., Narayana, S. V. L., DeLucas, L. J., El-Kabbani, O., Kilpatrick, J. M., Volanakis, J. E. & Bugg, C. E. (1991). *Proceedings of the Second Conference on Computer Visualization and Imaging*, Univ. of Iowa, pp. 29–42. Univ. of Iowa Press.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). *Nature (London)*, **342**, 877–883.

Fehlhammer, H., Bode, W. & Huber, R. (1977). *J. Mol. Biol.* **111**, 415–438

Feng, D. F. & Doolittle, R. F. (1987). *J. Mol. Evolution*, **25**, 351–360.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, N.-H. (1970). *Biochemistry*, **9**, 1997–2009.

Fujinaga, M. & James, M. N. G. (1987). *J. Mol. Biol.* **195**, 373–396.

Greer, J. (1988). Am. Crystallogr. Assoc. Meet., Abstract I3.

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

Hermans, J. & McQueen, J. E. (1974). *Acta Cryst.* A**30**, 730–739.

Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

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

Lesavre, P. & Muller-Eberhard, J. J. (1978). *J. Exp. Med.* **148**, 1498–1509.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Narayana, S. V. L., Carson, M., El-Kabbani, O., Kilpatrick, J. M., Moore, D., Chen, X., Bugg, C. E., Volanakis, J. E. & DeLucas, L. J. (1994). *J. Mol. Biol.* **235**, 695–708.

Narayana, S. V. L., Kilpatrick, J. M., El-Kabbani, O., Babu, Y. S., Bugg, C. E., Volanakis, J. E. & DeLucas, L. J. (1991). *J. Mol. Biol.* **219**, 1–3.

NAVIA, M. A., MCKEEVER, B. M., SPRINGER, J. P., LIN, T.-Y., WILLIAMS, H. R., FLUDER, E. M., DORN, C. P. & HOOGSTEEN, K. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 7–11.

PEARSON, W. R. & LIPMAN, D. J. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

PELLGRATH, J. W., SAPER, M. A., & QUIOCHO, F. A. (1984). *Methods and Applications in Crystallographic Computing*, edited by S. HALL & T. ASHIDA, pp. 404–407. Oxford: Clarendon Press.

PONDER, J. W. & RICHARDS, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.

REMINGTON, S. J., WOODBURY, R. G., REYNOLDS, R. A., MATTHEWS, B. W. & NEURATH, H. (1988). *Biochemistry*, **27**, 8097–8105.

ROSSMANN, M. G. (1972). *The Molecular Replacement Method*. New York: Gordon and Breach.

SAWYER, L., SHOTTON, D. M., CAMPBELL, J. W., WENDELL, P. L., MUIRHEAD, H., WATSON, H. C., DIAMOND, R. & LADNER, R. C. (1978). *J. Mol. Biol.* **118**, 137–208.

SIBANDA, B. L., BLUNDELL, T., HOGART, P. M., FOGLIANO, M., BINDRA, J. S., DOMINY, B. W. & CHIRGWIN, (1984). *FEBS Lett.* **174**, 102–111.

SPRANG, S., STANDING, T., FLETTERICK, R. J., STROUD, R. M., FINER-MOORE, J., XUONG, N.-H., HAMLIN, R., RUTTER, W. J. & CRAIK, C. S. (1987). *Science*, **237**, 905–909.

VIJAY-KUMAR, S. & COOK, W. J. (1992). *J. Mol. Biol.* **224**, 413–426.