International Tables for Crystallography (2006). Vol. F, Section 21.3.5, pp. 521–525.

The *PROCHECK* suite is generally useful for assessing the quality of protein structures in various stages of completion. The Ramachandran analysis is especially informative. However, it is possible, at least in principle, to devise an incorrect model with fully acceptable main-chain and side-chain stereochemistry, so other methods must also be used to assess protein models.

21.3.3.2. *WHAT IF*

The molecular modelling and drug design program *WHAT IF* (Vriend, 1990) performs a large number of geometrical checks, comparing a proposed protein model to a set of canonical distances and angles. These parameters include bond lengths and bond angles, side-chain planarity, torsion angles, interatomic distances, unusual backbone conformations and the Ramachandran plot. New additions (Vriend & Sander, 1993) include a 'quality factor', and a number of checks for clashes between symmetry-related molecules. Starting from the hypothesis that atom–atom interactions are the primary determinant of protein folding, the program tests a protein model for proper packing by calculating a contact quality index. Each contact is characterized by its fragment type (80 types from the 20 residues), the atom type and the three-dimensional location of the atom relative to the local frame of the fragment. Sets of database-derived distributions are compared with the actual distribution in the protein model being tested. A good agreement with the database distribution produces a high contact quality index. A low packing score can indicate any of: poor packing, misthreading of the sequence, bad crystal contacts, bad contacts with a co-factor, or proximity to a vacant active site. The contact analysis available in *WHAT IF* can be used as an independent quality indicator during crystallographic refinement, or during the process of protein modelling and design.

21.3.3.3. *VERIFY3D*

The program *VERIFY3D* (Lüthy *et al.*, 1992; Bowie *et al.*, 1991) measures the compatibility of a protein model with its own amino-acid sequence. Each residue position in the 3D model is characterized by its environment and is represented by a row of 20 numbers in a '3D profile'. These numbers are the statistical preferences, called 3D–1D scores, of each of the 20 amino acids for this environment. Environments of residues are defined by three parameters: the area of the residue buried in the protein and inaccessible to solvent, the fraction of the side-chain area that is covered by polar atoms (O and N), and the local secondary structure. The 3D profile score, *S*, for the compatibility of the sequence with the model is the sum, over all residue positions, of the 3D–1D scores for the amino-acid sequence of the protein. The compatibility of segments of the sequence with their 3D structures can be assessed by plotting, against sequence number, the average 3D–1D score in a window of 21 residues. The 3D profile method rests on the observation that soluble proteins bury many hydrophobic side chains and not many polar residues.

Three applications for 3D profiles exist. The first is to assess the validity of protein models (Lüthy *et al.*, 1992). For 3D protein models known to be correct, the 3D profile score, *S*, for the compatibility of the amino-acid sequence with the environments formed by the model is high. In contrast, *S* for the compatibility with its sequence of a totally or partially wrong 3D protein model is generally low. Therefore, models that are largely incorrect or models that contain a small number of improperly built segments can be detected by low-scoring regions in the 3D profile. However, not all faulty regions are always evident directly from the profile, particularly if the misbuilt regions are at the termini, where they are obscured by the windowing procedure. The second application is to assess which is the stable oligomeric state of the folded protein, by comparing the accessibility (buried or exposed) of amino-acid side chains in the monomeric and oligomeric state (Eisenberg *et al.*, 1992). The third application is to identify other protein sequences which are folded in the same general pattern as the structure from which the profile was prepared (Bowie *et al.*, 1991). Predicting a protein structure from sequence requires a link between 3D structure and 1D sequence. The program *VERIFY*3*D* provides this link by reducing a 3D structure to 1D string of environmental classes. Therefore the method can be used to evaluate any protein model or to measure the compatibility of any protein structure with its amino-acid sequence.

21.3.3.4. *ERRAT*

The program *ERRAT* (Colovos & Yeates, 1993) analyses the relative frequencies of noncovalent interactions between atoms of various types. It can be viewed as an extension of the earlier 3D profile approach from the residue level to the atom level. Three types of atoms are considered (C, N and O), and consequently six types of interactions are possible (CC, CN, CO, NN, NO and OO).

ERRAT operates under the hypothesis that different atom types will be distributed non-randomly with respect to each other in proteins due to complex geometric and energetic considerations, and that structural errors will lead to detectable anomalies in the pattern of interactions. Assessment of the non-bonded interactions is subject to the following restrictions: the distance between the two atoms in space is less than some preset limit, typically 3.5 Å, and the atoms within the same residue or those that are covalently bonded to each other are not considered. For each nine-residue segment of sequence, the non-bonded contacts to other atoms in the protein are tallied by atomic interaction type and the result is divided by the total number of interactions. This gives a list, or six-dimensional vector, of fractional interaction frequencies that add up to unity. In this way, each nine-residue fragment generates one point in a five-dimensional space; only five of the six fractional values are independent. A large number of observations were extracted from reliable high-resolution structures and used to establish a multivariate five-dimensional normal distribution for accurate protein structures. This distribution is used to evaluate the probability that a given set of interactions from a protein model in question is correct. Since the *ERRAT* evaluation is based on a normal distribution calibrated on a reliable database, it is straightforward to estimate the likelihood that each region of a candidate protein model is incorrect. This method provides an unbiased and statistically sound tool for identifying incorrectly built regions in protein models.

21.3.4. Selection of database

Regardless of the specific approach or the specific criteria for validating structural models, a reliable reference database has to be chosen by careful selection of known structures. Suitable criteria to consider when selecting a database are: protein structures determined to resolutions of 2.5 Å or better, *R* factors less than 25%, and good geometry, particularly of the dihedral angles of the protein backbone. In addition, the database should include examples from many diverse classes of structures and at the same time avoid multiple identical structures.

21.3.5. Examples: detection of errors in structures

21.3.5.1. *Specific examples*

Several examples are presented of errors in structural models determined by X-ray crystallography that can be detected using validation methods. One is that of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), which was traced essentially backwards from a poor electron-density map (Chapman *et al.*, 1988). The program *ERRAT* finds that approximately 40% of the residues in this mistraced model are outside the 95% confidence limit (Fig. 21.3.5.1*a*). This limit is the error value above which a given region can be judged to be erroneous with 95% certainty, so a reliable model should exceed this value over less than 5% of its length. The final model of RuBisCO (Curmi *et al.*, 1992) shows only 2% of the residues outside *ERRAT*'s 95% confidence limit. Similarly, the 3D profile calculated from *VERIFY3D* for the erroneous model (Fig. 21.3.5.1*b*) gives a total score of 15 when matched to the sequence of the small subunit of RuBisCO. This score is well below the expected value of 58 for the correct structure of this length. Indeed, the 3D profile of the correct model (Curmi *et al.*, 1992) (Fig. 21.3.5.1*b*) of RuBisCO has a score of 55. *PROCHECK* and *WHAT IF* also identify stereochemistry problems in the original model, including deviant bond angles and bond lengths, many residues in the disallowed Ramachandran regions (Fig. 21.3.5.1*c*), bad peptide-bond planarity, and bad non-bonded interactions. In contrast, most amino-acid residues of the correct RuBisCO model are in the allowed regions of the Ramachandran plot (Fig. 21.3.5.1*d*) with good overall geometry.

The archive of obsolete PDB entries maintained by the San Diego Supercomputer group (http://pdbobs.sdsc.edu) includes old versions of protein structures that have been withdrawn and/or replaced by the depositor with a newer version. One example is that of a protein (3xia.coor) originally solved to 3 Å in the wrong space group and later to 1.8 Å in the correct space group (1xya.coor). The *ERRAT* program reveals problems in the original model, with 45% of the residues outside the 95% confidence limit (Fig. 21.3.5.2*a*). The more recent model has only 1.5% of the residues outside the

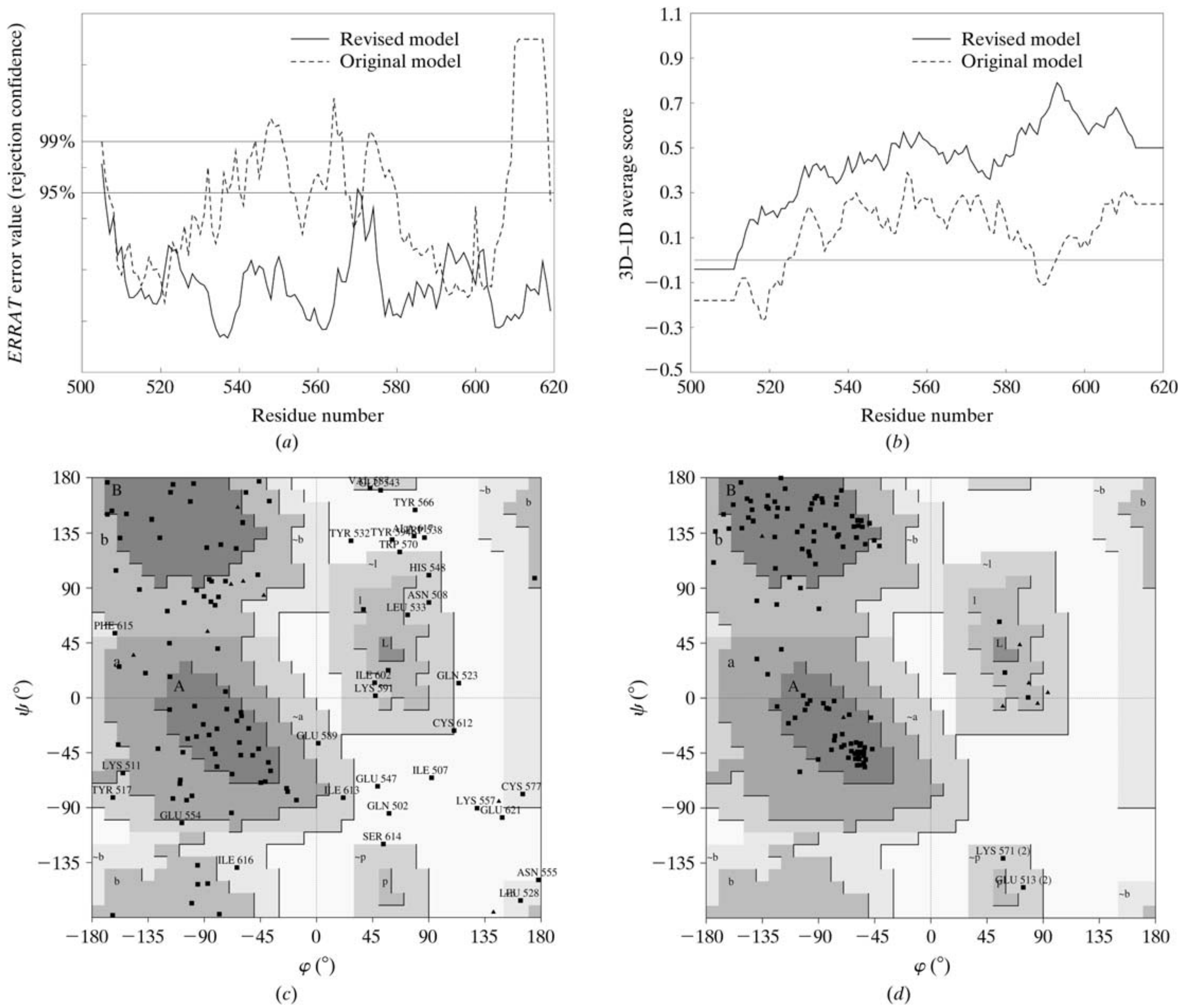

Fig. 21.3.5.1. Detection of errors in the small subunit of ribulose-1,5-bisphospate carboxylase/oxygenase (RuBisCO). (*a*) *ERRAT* plot of the error function in a nine-residue sliding window, the centre of which is at the sequence position indicated by the horizontal axis. The solid bold line represents the revised structure and the dashed line the original structure. The thin solid lines indicate the 95% and 99% confidence limits for rejection. A region above the 95% line can be judged incorrect with 95% certainty. (*b*) *VERIFY3D* profile-window plots for the revised (bold) and original (dashed) models. The vertical axis gives the average 3D–1D score for residues within a 21-residue sliding window. Regions that score below zero are suspect. (*c*) Ramachandran diagram from *PROCHECK* of the initial structure of RuBisCO. The main-chain dihedral angle φ (N—Cα bond) is plotted *versus* ψ (Cα—C bond). All non-glycine residues outside the allowed regions are marked. (*d*) Ramachandran plot for the refined RuBisCO structure.

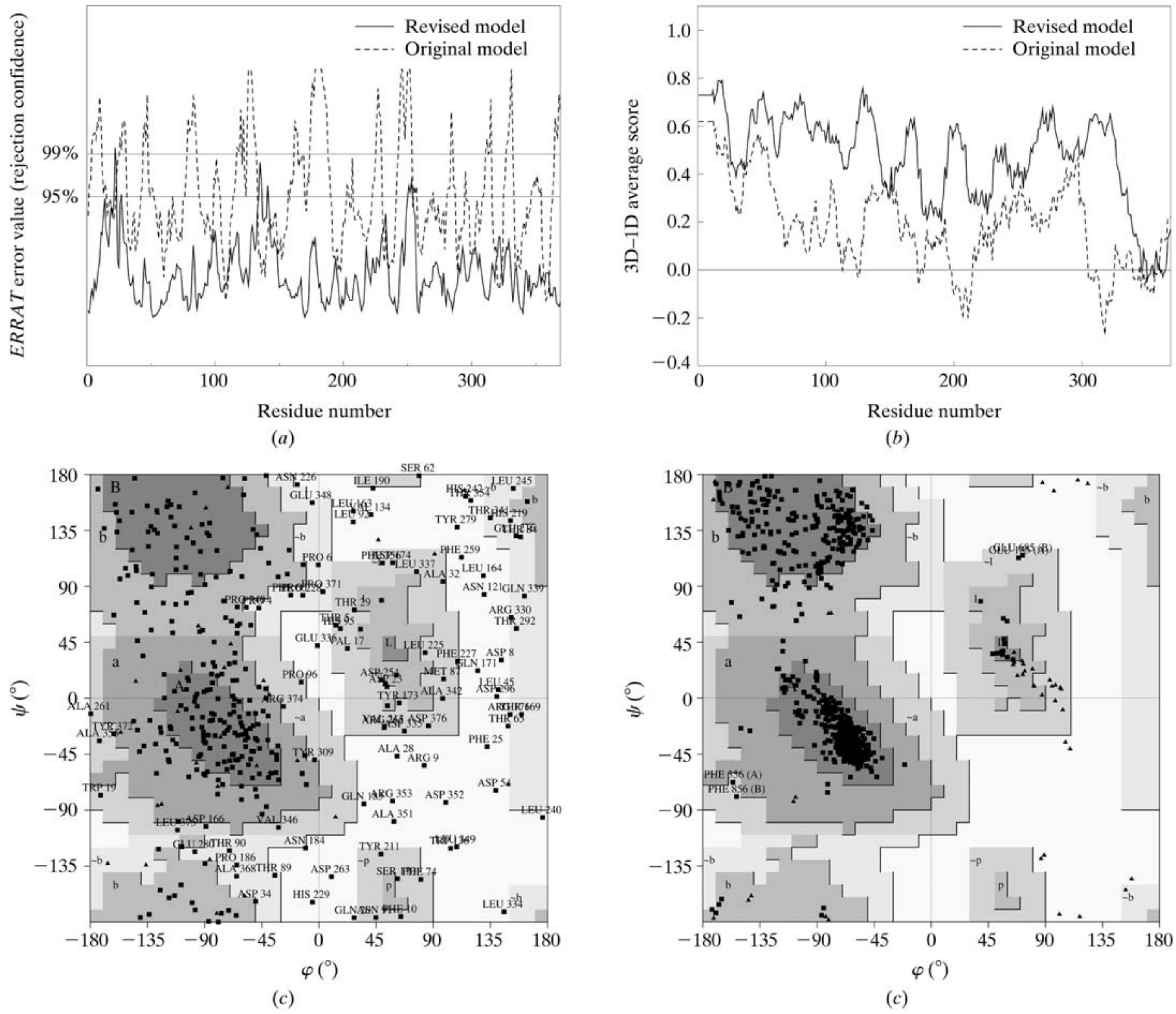

Fig. 21.3.5.2. The detection of model errors due to refinement in an incorrect space group: an example (3xia.coor) from the archive of obsolete PDB entries. (*a*) *ERRAT* plot of the error function in a nine-residue sliding window. The solid bold line represents the revised structure and the dashed line represents the original structure. The thin solid lines indicate the 95% and 99% confidence limits for rejection. (*b*) *VERIFY3D* profile-window plots for the revised (bold) and original (dashed) models. The vertical axis gives the average 3D–1D score for residues within a 21-residue sliding window. (*c*) Ramachandran diagram of the original structure. All non-glycine residues outside the allowed regions are marked. (*d*) Ramachandran diagram for the revised structure.

95% confidence limit. The problem in the original model is also illustrated by the *VERIFY3D* plot (Fig. 21.3.5.2*b*) for which the average score is often below the value of 0.1 and dips below zero at four points. In contrast, the *VERIFY3D* plot of the revised model shows no dips below zero. Poor stereochemistry is also apparent in the Ramachandran plot of the original model (Fig. 21.3.5.2*c*). Only 38% of the backbone dihedral angles lie in the most favoured regions, compared to 93.8% in the revised model (Fig. 21.3.5.2*d*).

The potential usefulness of error-detecting programs during model building is suggested by stages in the crystal structure determination of triacylglycerol lipase from *Pseudomonas cepacia* (Kim *et al.*, 1997), which was solved by MIR. The authors kindly provided us with ten different models (assigned as stage number 1–10) along the course of model building and refinement. Regions where Cα positions shifted between initial and final models correlated with regions where the error functions improved. For example, the program *ERRAT* points at specific regions (*e.g.* 18–35 and 135–165) originally assigned as polyalanine. When at the next stage of refinement these were changed to the actual amino-acid sequence, these regions behaved normally (Fig. 21.3.5.3*a*). This illustrates that *ERRAT* is able to illuminate problem areas in a structure.

VERIFY3D is sensitive to unusual environments in proteins. An illustration is offered by the structures of lipases, with and without their inhibitors. There are two general conformations known as ‘closed’ and ‘open’. In the so-called ‘closed’ structure, the catalytic triad is buried underneath a helical segment, called a ‘lid’ (Brzozowski *et al.*, 1991), so that hydrophobic residues tend to be buried as observed in a ‘normal’ 3D profile. In the ‘open’ conformation, the lipid binding site becomes accessible to the solvent, and hydrophobic surfaces (residues 140–150 and 230–250) are exposed by the movement of the ‘lid’. These hydrophobic

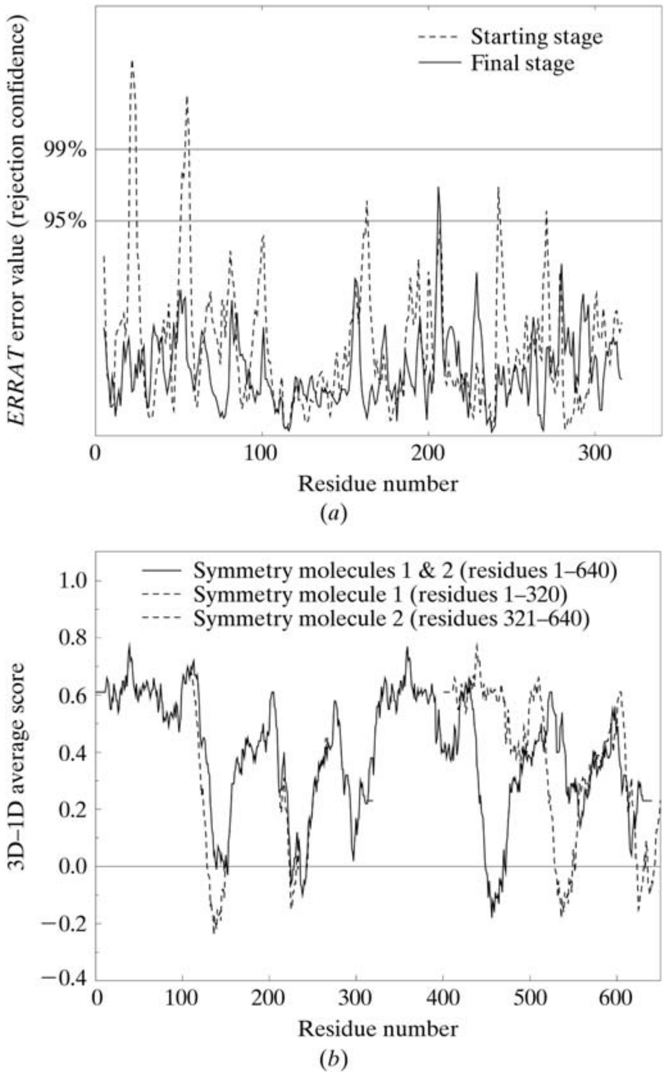

Fig. 21.3.5.3. The usefulness of validation programs during model building is suggested by the example of the triacylglycerol lipase from *Pseudomonas cepacia* at different stages of atomic refinement. (*a*) Plot from *ERRAT* at the initial and final stages of refinement. (*b*) *VERIFY3D* profile-window plots of the final model. The dashed line represents symmetry molecule number 1 (residues 1–320) and symmetry molecule number 2 (residues 321–640) when not in contact with each other. The solid line represents symmetry molecule 1 and 2 when in contact. This plot illustrates that the state of oligomerization can affect the 3D profile plot, giving information on the oligomerization. See Bennett *et al.* (1994) for more details.

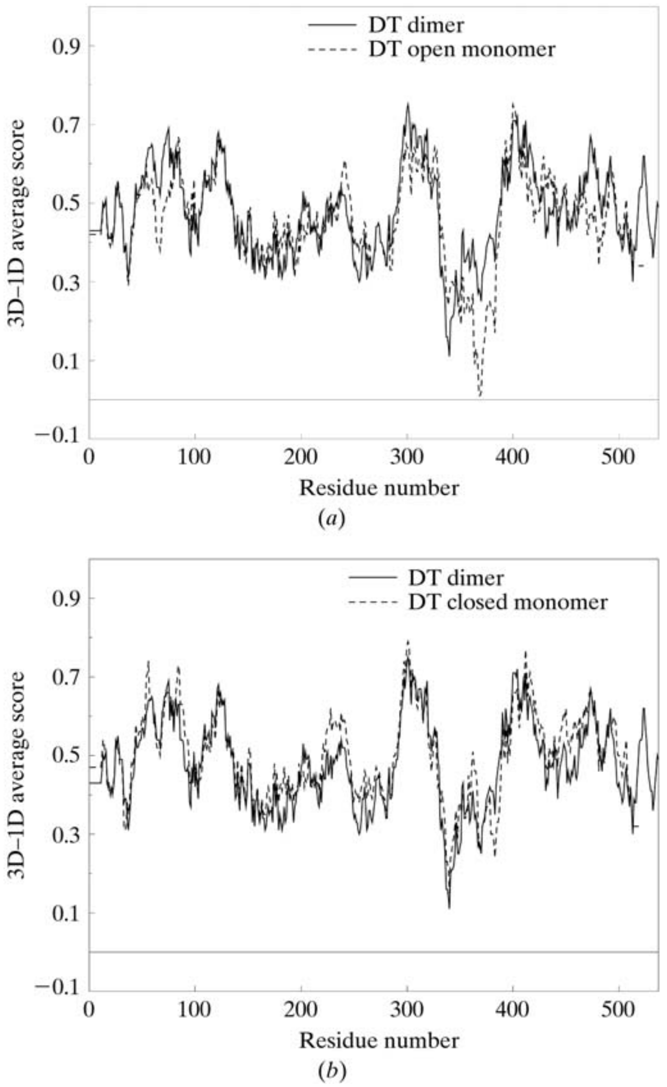

Fig. 21.3.5.4. *VERIFY*3D profile plots of diphtheria toxin (DT) in three forms: open and closed monomers and the dimer. (*a*) DT open monomer (dashed), DT dimer (solid line). (*b*) DT closed monomer (dashed) and dimer (solid line). Notice that the hinge loop (residues 379–387) in the open monomer has a low profile score, and this structure is known to be unstable. The score is raised in the stable closed monomer and in the dimer.

exposed regions are strikingly shown in the 3D profile of the 'open' structure (Fig. 21.3.5.3*b*), which clearly reveals the two problematic regions (140–150 and 230–250) with profile scores below zero. The exposed hydrophobic residues 140–150 from one symmetry model make van der Waals interactions with hydrophobic residues 230–250 from a symmetry-related molecule (Kim *et al.*, 1997). These interactions are revealed as higher scores in those regions when inspecting the 3D profiles of the two symmetry-related molecules.

Another example of unusual environment is that of diphtheria toxin (DT), which exists as a monomer as well as a dimer. Monomeric DT is a Y-shaped molecule with three domains known as catalytic (C), transmembrane (T) and receptor binding domain (R). Crystal structures have been determined for both the 'closed' monomeric form and for a domain-swapped dimeric form (Bennett *et al.*, 1994). Upon dimerization, a massive conformational rearrangement occurs and the entire R domain from each monomer of the dimer is interchanged with the other monomer. This involves breaking the noncovalent interactions between the R domain and the C and T domains and rotating the R domain by 180° with atomic movements up to 65 Å to produce the 'open' conformation. After rearrangement, each R domain reforms the same noncovalent interactions as it had in the monomer, but with the C and T domains of the other monomer. The existence of both open and closed forms of DT requires that large conformational changes occur in residues 379–387 (the hinge loop). The 3D profile of the 'open' form (Fig. 21.3.5.4*a*) shows low scores for these residues compared to the closed monomer or dimer (Fig. 21.3.5.4*b*). The higher scores of the open monomer are consistent with the greater stability of the monomer in the closed rather than the open conformation.

21.3.5.2. *Survey of old and revised structures*

The past two decades have seen a surge of development in the experimental techniques of crystal structure determination. As a consequence, many structures originally solved at low resolution were later determined at higher resolution, often starting with improved phases. The archive of obsolete PDB entries maintained

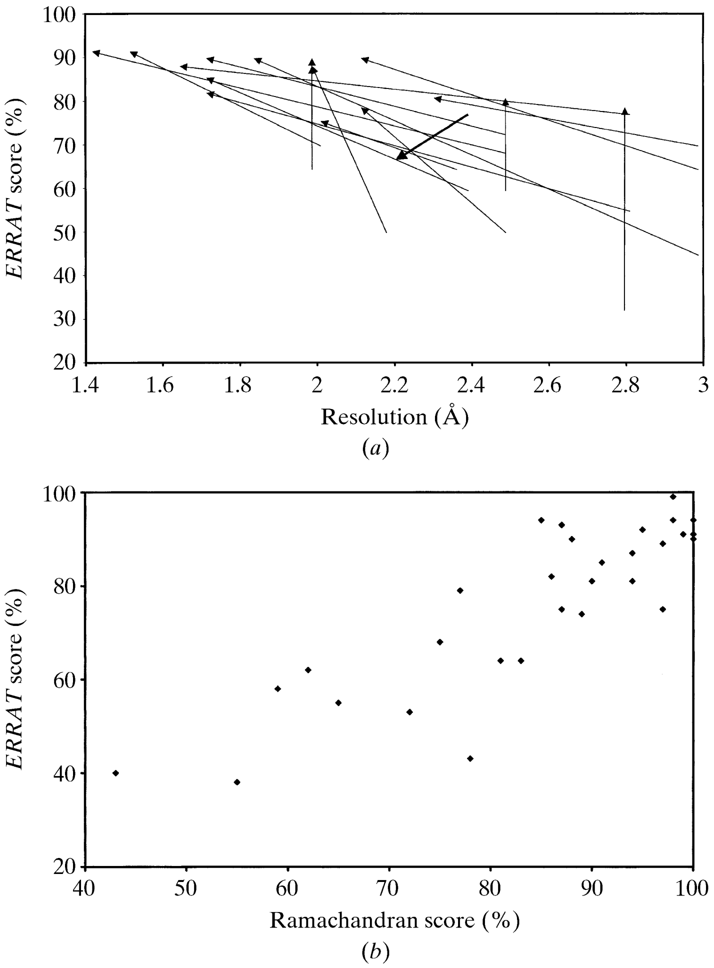

Fig. 21.3.5.5. Evaluation of old and revised models in a database survey by *ERRAT*. (*a*) Percentage of residues within the 95% confidence limit given by *ERRAT* as a function of the resolution to which the structure was determined. Each arrow represents an obsolete structure (arrow tail) and the revised structure that replaced it (arrow head). The revised structures (typically analysed at finer resolution) show markedly improved *ERRAT* scores. (*b*) Correlation between the percentage of a structure within the 95% confidence limit according to *ERRAT*, and the percentage of residues in the most favoured regions of a Ramachandran diagram according to *PROCHECK*.

by the San Diego Supercomputer group (http://pdbobs.sdsc.edu) served as a benchmark for evaluating the *ERRAT* program. For testing, 17 pairs of protein models were selected. Each pair comprised an obsolete entry and the revised model that replaced it. Using *ERRAT*, the overall quality of each model was expressed as a single number according to the fraction of the structure falling below the 95% confidence limit for rejection. The overall scores are significantly better for the revised structures, most of which were analysed at improved resolution (Fig. 21.3.5.5*a*). This result further demonstrates the utility of *ERRAT* for monitoring the model-building process. Furthermore, a strong correlation is found between the percentage of residues within the 95% confidence limit given by *ERRAT* and the percentage of residues in the most favoured regions of the Ramachandran plot of *PROCHECK* (Fig. 21.3.5.5*b*). In general, the problematic regions detected by the two programs agree with each other.

21.3.6. Summary

In order to ensure the quality of the growing protein structure databases, models must be evaluated carefully during and after the structure determination process. Model evaluation can incorporate two types of measures: agreement between the model and the experimental diffraction data, and agreement between the model and the database of known structures. The latter types use the atomic coordinates of the final model, but do not rely on the diffraction data. In recent years, powerful methods of this type have been developed.

The most informative and reliable model-evaluation criteria are those that measure properties not optimized as part of the automatic refinement procedure. The free *R* value has become important for monitoring the progress of atomic refinement for the same reason: it is based on reflections not included in refinement. We have focused here on two programs, *VERIFY3D* and *ERRAT*, which both evaluate high-level geometric properties not optimized during atomic refinement. Each offers the convenience of a single score over a sliding window along the protein sequence. Because *VERIFY3D* operates on the level of amino-acid residues, it is sensitive to errors on that scale, particularly those that affect the distribution of polar and nonpolar residues. *ERRAT* operates on the atomic level and has proven to be particularly useful for pinpointing local regions of protein models that require further adjustments. When used in combination, these methods and others can help crystallographers produce more accurate structural models of proteins.

21.3.7. Availability of software

The programs *ERRAT* and *VERIFY3D* are available on the World Wide Web for non-commercial applications. The URL for *VERIFY3D* is http://www.doe-mbi.ucla.edu/Services/Verify3D.html and the URL for *ERRAT* is http://www.doe-mbi.ucla.edu/Services/Errat.html. *VERIFY3D* and *ERRAT* expect a coordinate file in PDB format. The programs return plots of the type shown in this chapter.

Acknowledgements

We thank Dr Kyeong Kyu Kim and Dr Se Won Suh of the Department of Chemistry, Seoul National University, Korea, for the models of triacylglycerol lipase. This work was supported by grants NIH GM 31299, DOE DE-FG03-87ER60615 and NSF MCB 9420769.

references