25.2.4.6.3. *Randy Read's maximum-likelihood function*

When Navraj Pannu wanted to implement Read's maximum-likelihood refinement functions (Pannu & Read, 1996*b*) in *TNT*, he choose not to implement it as a separate program, but modified *TNT*'s source code to create a new version of the program *Rfactor*, named *Maxfactor*.

25.2.4.6.4. *J. P. Abrahams' likelihood-weighted noncrystallographic symmetry restraints*

Abrahams (1996) conceived the idea that because some amino-acid side chains can be expected to violate the noncrystallographic symmetry (NCS) of the crystal more than others, one could develop a library of the relative strength with which each atom of each residue type would be held by the NCS restraint. He chose to determine these strengths from the average of the current agreement to the NCS of all residues of the same type. For example, if the lysine side chains do not agree well with their NCS mates, the NCS will be loosely enforced for those side chains. On the other hand, if almost all the valine side chains agree well with their mates, then the NCS will be strongly enforced for the few that do not agree well.

He chose to implement this idea by modifying the source code for the *TNT* program *NCS*. Since the calculations involved in implementing this idea are simple, the extent of the modifications were not large.

25.2.5. The *ARP/wARP* suite for automated construction and refinement of protein models (V. S. Lamzin, A. Perrakis and K. S. Wilson)

25.2.5.1. *Refinement and model building are two sides of modelling a structure*

The conventional view of crystallographic refinement of macromolecules is the optimization of the parameters of a model to fit both the experimental data and a set of *a priori* stereochemical observations. The user provides the model and, although the values of its parameters are allowed to vary during the minimization cycles, the presence of the atoms is fixed, *i.e.* the addition or removal of parts of the model is not allowed. As a result, users are often faced with a situation where several atoms lie in one place, while the density maps suggest an entirely different location. Manual intervention, consisting of moving atoms to a more appropriate place using molecular graphics, density maps and geometrical assumptions can solve the problem and allow refinement to proceed further.

The *Automated Refinement Procedure* (*ARP*; Fig. 25.2.5.1) (Lamzin & Wilson, 1993, 1997; Perrakis *et al.*, 1999) challenges this classical view by addition of *real-space manipulation* of the model, mimicking user intervention *in silica.* Adding and/or deleting atoms (*model update*) and complete re-evaluation of the model to create a new one that better describes the electron density (*model reconstruction*) can achieve this aim.

25.2.5.1.1. *Model update*

The quickest way to change the position of an atom substantially is not to move it, but rather involves a two-step procedure to remove it from its current (probably wrong) site and to add a new atom at a new (hopefully right) position. Such updating of the model does not imply that all rejected atoms are immediately repositioned in a new site, so the number of atoms to be added does not have to be equal to the number rejected.

Atom rejection in *ARP* is primarily based on the interpolated $2mF_o - \Delta F_c$ or $3F_o - 2F_c$ electron density at its atomic centre and the agreement of the atomic density distribution with a target shape.

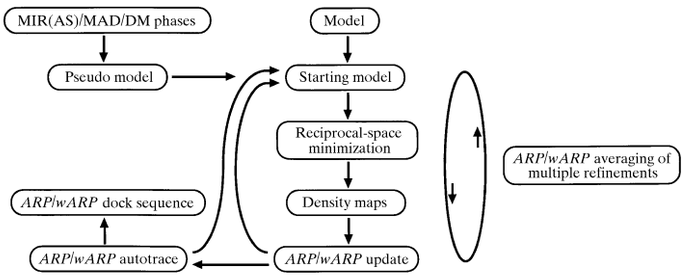

Fig. 25.2.5.1. A flow chart of the *Automated Refinement Procedure*.

Applied together, these criteria offer powerful means of identifying incorrectly placed atoms, but can suggest false positives. However, a correctly located atom that happens to be rejected should be selected again and put back in the model. Developments of further, perhaps more elegant, criteria may be expected in the future development of the technique.

Atom addition uses the difference $mF_o - \Delta F_c$ or $F_o - F_c$ Fourier synthesis. The selection is based on grid points rather than peaks, as the latter are often poorly defined and may overlap with neighbouring peaks or existing atoms, especially if the resolution and phases are poor. The map grid point with the highest electron density satisfying the defined distance constraints is selected as a new atom, grid points within a defined radius around this atom are rejected and the next highest grid point is selected. This is iterated until the desired number of new atoms is found and reciprocal-space minimization is used to optimize the new atomic parameters.

Real-space refinement based on density shape analysis around an atom can be used for the definition of the optimum atomic position. Atoms are moved to the centre of the peak using a target function that differs from that employed in reciprocal-space minimization. The function used is the sphericity of the site, which keeps an atom in the centre of the density cloud but has little influence on the *R* factor and phase quality. It is only applicable for well separated atoms and is mainly used for solvent atoms at high resolution.

Geometrical constraints are based on *a priori* chemical knowledge of the distances between covalently linked carbon, nitrogen and oxygen atoms (1.2 to 1.6 Å) and hydrogen-bonded atoms (2.2 to 3.3 Å). Such constraints are applied in rejection and addition of atoms.

25.2.5.1.2. *Model reconstruction*

The main problem in automatically reconstructing a protein model from electron-density maps is in achieving an initial tracing of the polypeptide chain, even if the result is only partially complete. Subsequent building of side chains and filling of possible gaps is a relatively straightforward task. The complexity of the autotracing can be nicely illustrated as the well known travelling-salesman problem. Suppose one is faced with 100 trial peptide units possessing two incoming and two outgoing connections on average, which is close to what happens in a typical *ARP* refinement of a 10 kDa protein. Assuming that one of the chain ends is known and that it is possible to connect all the points regardless of the chosen route, then one is faced with the problem of choosing the best chain out of 2^{98}. In practice, the situation is even more complex, as not all trial peptides are necessarily correctly identified in the first iteration and some may be missing – analogous to the correctness or incorrectness of the atomic positions described above.

If the connections can be assigned a probability of the peptide being correct, then only the path that visits each node exactly once and maximizes the total probability remains to be identified. Automatic density-map interpretation is based on the location of the atoms in the current model and consists of several steps. Firstly,

each atom of the free-atom model is assigned a probability of being correct. Secondly, these weighted atoms are used for identification of patterns typical for a protein. The method utilizes the fact that all residues that comprise a protein, with the exception of *cis* peptides, have chemically identical main-chain fragments which are close to planar: the structurally identical Cα—C—O—N—Cα *trans* peptide units.

The problem of searching for possible peptide units and their connections thus becomes straightforward. The most crucial factor is that proteins are composed of linear non-branching polypeptide chains, allowing sets of connected peptides to be obtained from an initial list of all possible tracings. Choosing the direction of a chain path is carried out on the basis of the electron density and observed backbone conformations. The set of peptide units and the list of how they are interconnected do not, however, allow unambiguous tracing of a full-length chain in most cases.

Taken together, the probabilistic identification of the peptide units, the naturally high conformational flexibility of the connections of the peptide units and the limited quality of the X-ray data and/or phases introduce large enough errors to cause density breaks in the middle of the chains or result in density overlaps. Thus, the result of such a tracing is usually a set of several main-chain fragments. The less accurate the starting maps (*i.e.* initial phases) and the lower the resolution and quality of the X-ray data, the more breaks there will be in the tracing and the greater the number of peptide units which will be difficult to identify.

Residues are differentiated only as glycine, alanine, serine and valine, and complete side chains are not built at this stage. For every polypeptide fragment, a side-chain type can be assigned with a defined probability, using connectivity criteria from the free-atom models and the α-carbon positions of the main-chain fragments. Given these guesses for the side chains and provided the sequence is known, the next step employs docking of the polypeptide fragments into the sequence. Each possible docking position is assigned a score, which allows automated inspection of the side-chain densities, search for expected patterns and building of the most probable side-chain conformations.

25.2.5.1.3. *Representation of a map by free-atom models*

An electron-density map can be used to create a free-atom atomic model, with equal atoms placed in regions of high density (Perrakis *et al.*, 1997). To build this model, only the molecular weight of the protein is required, without any sequence information. In brief, a map covering a crystallographic asymmetric unit on a fine grid of about 0.25 Å is constructed. The model is slowly expanded from a random seed by the stepwise addition of atoms in significant electron density and at bonding distances from existing atoms. All atoms in this model and in all subsequent steps are considered to be of the same type. As *ARP* proceeds, the geometrical criteria remain the same, but the density threshold is gradually reduced, allowing positioning of atoms in lower-density areas of the map. The procedure continues until the number of atoms is about three times that expected. This number is then reduced to about $n + 20\%$ atoms by removing atoms in weak density. This method of map parameterization has the advantage that it puts atoms at protein-like distances while covering the whole volume of the protein.

25.2.5.1.4. *Hybrid models*

A free-atom model can describe almost every feature of an electron-density map, but this interpretation rarely resembles a conventional conception of a protein. Nevertheless, information from parts of the improved map and the free-atom model can be automatically recognized as containing elements of protein structure by applying the algorithms briefly described for model reconstruction, and at least a partial atomic protein model can be built. Combination of this partial protein model with a free-atom set (a *hybrid* model) allows a considerably better description of the current map. The protein model provides additional information (in the form of stereochemical restraints), while prominent features in the electron density (unaccounted for by the current model) are described by free atoms.

25.2.5.1.5. *Real-space manipulation coupled with reciprocal-space refinement*

The procedure of real-space manipulation is coupled to least-squares or maximum-likelihood optimization of the model's parameters against the X-ray data. This is the scheme that we generally refer to as *ARP* refinement, though there are two distinct modes of *ARP*: In the *unrestrained* mode, all atoms in reciprocal-space refinement are treated as free atoms with unknown connectivity and are refined against the experimental data alone. This mode has a higher radius of convergence but needs high-resolution diffraction data to perform effectively. In the *restrained* mode, a model or a hybrid model is required, *i.e.* the atoms must belong to groups of known stereochemistry. This stereochemical information, in the form of restraints, can then be utilized during the reciprocal-space minimization, allowing it to proceed with less data, presuming that the connectivity of the input atoms is basically correct.

25.2.5.2. *ARP/wARP applications*

25.2.5.2.1. *Model building from initial phases*

The hybrid models described above are used as the main tool for obtaining as full a protein model as possible from the map calculated with the initial phases.

Given the information contained in the hybrid model in the traditional form of stereochemical restraints, reciprocal-space refinement can work more efficiently, new improved phases can be obtained and a more accurate and complete protein model can be constructed. The new hybrid model can be re-input to refinement and these steps can be iterated so that improved phases result in construction of ever larger parts of the protein. An almost complete protein model can be obtained in a fully automated way.

25.2.5.2.2. *Refinement of molecular-replacement solutions*

Starting from a molecular-replacement solution implies that a search model positioned in the new lattice is already available. The model can be directly incorporated in restrained *ARP* refinement. If the starting model is very incomplete or different, its atoms can be regarded as free atoms and the solution can be treated as starting from just initial phases. This increases the radius of convergence and minimizes the bias introduced by the search model.

25.2.5.2.3. *Density modification via averaging of multiple refinements*

Slightly varying the protocol described for generating models from maps results in a set of slightly different free-atom models. Each model is then submitted to *ARP*. In protein crystallography, there are generally insufficient data for convergence of free-atom refinement to a global minimum and different starting models result in final models with small differences, *i.e.* containing different errors. Averaging of these models can be utilized to minimize the overall error. The procedure in effect imposes a random noise, small enough to be eliminated during the subsequent averaging, but large enough to overcome at least some of the systematic errors.

Structure factors are calculated for all the refined models and a vector average of the calculated structure factors is derived. The phase of the vector average is more accurate than that from any of

the individual models. A weight, W_{wARP}, is assigned to each structure factor on the basis of the variance of the two-dimensional distribution of the individual structure factors around the mean. The mean value of W_{wARP} over all reflections and the R factor after averaging can be used to judge the progress of the averaging procedure.

25.2.5.2.4. *Ab initio solution of metalloproteins*

If the coordinates of one or a few heavy atoms are known, initial phases can be calculated. The problem of solving the structure of such a metalloprotein from the sites of the metal alone can be considered in the same framework as for heavy-atom-replacement solutions. Maps calculated from the phases of heavy atoms alone often have the best defined features within a defined radius of the heavy atom(s). Thus protocols that do not place all atoms at the start but instead perform a slow building while extending the model in a growing sphere around the heavy atom are preferred. When such a model is essentially complete, it can be used for automated tracing and completion of the model.

25.2.5.2.5. *Solvent building*

In this application, the protein (or nucleic acid) model is not rebuilt during refinement, and only the solvent structure is continuously updated, allowing the construction of a solvent model without iterative manual map inspection.

25.2.5.3. *Applicability and requirements*

Density-based atom selection for the whole structure is only possible if the X-ray data extend to a resolution where atomic positions can be estimated from the Fourier syntheses with sufficient accuracy for them to refine to the correct position. If the structural model is of reasonable quality, at 2.5 Å or better, at least a part of the solvent structure or a small missing or badly placed part of the protein can be located. This provides indirect improvement of the whole structure. For automated model rebuilding, or for refining poor molecular-replacement solutions, higher resolution is essential. The general requirement is that the number of X-ray reflections should be at least six to eight times higher than the number of atoms in the model, which roughly corresponds to a resolution of 2.3 Å for a crystal with 50% solvent. However, the method can work at lower resolution or fail with a higher one, depending less on the quality of the initial phases and more on the internal quality of the data and on the inherent disorder of the molecule.

The X-ray data should be complete. If strong low-resolution data (*e.g.* 4 to 10 Å) are systematically missing, *e.g.* due to detector saturation, the electron density even for good models is often discontinuous. Because *ARP* involves updating on the basis of density maps, such discontinuity will lead to incorrect interpretation of the density and slow convergence or even uninterpretable output.

25.2.5.4. *An example*

The structure of chitinase A from *Serratia marcescens* (Perrakis *et al.*, 1994) was initially solved by multiple isomorphous replacement with anomalous signal (MIRAS), with only a single derivative contributing to resolution higher than 5.0 Å. The MIRAS map (2.5 Å) was solvent-flattened. Model building was not straightforward and much time was spent in tracing the protein chain.

As an experiment, the solvent-flattened map was used to initiate building of free-atom models, using least-squares minimization against the native 2.3 Å data combined with *ARP*. This resulted in crystallographic R factors ranging between 20.1 and 22.4%. Each *ARP* model gave phases marginally worse than those available by solvent flattening alone, due to the limited resolution of the native data. However, the *wARP* averaging procedure resulted in a reduction of 11.2° in the weighted mean phase error. The map correlation coefficient between the final map and the *wARP* map was 81.2%, better by 12.8% compared with the solvent-flattened map.

The *wARP* model with the lowest R factor was used to initiate model building. In the initial tracing, 75 residues were identified, belonging to more than 20 different main-chain fragments. After autobuilding, ten cycles of restrained *ARP* were run according to the standard protocol. One *REFMAC* cycle of conjugate-gradient minimization was executed to optimize a maximum-likelihood residual and bulk solvent scaling. σ_A-weighted maps were calculated and *ARP* was used to update the model. All atoms (main-chain, side-chain and free atoms) were allowed to be removed and new atoms were added where appropriate. After ten iterations, a new building cycle was invoked. After every 'big' cycle, a more complete model was obtained. This 'big' cycle was iterated 20 times. Finally, 515 residues were traced in nine chains, all of which were docked unambiguously into the sequence. This is the lowest-resolution application to date. 2.3 Å was the real resolution limit of the data measured from these crystals; however, the high solvent content (61%) provided on average seven observations per atom and an almost complete trace was easily accomplished.

25.2.6. *PROCHECK*: validation of protein-structure coordinates (R. A. LASKOWSKI, M. W. MACARTHUR AND J. M. THORNTON)

25.2.6.1. *Introduction*

As in all scientific measurements, the parameters that result from a macromolecular structure determination by X-ray crystallography (*e.g.* atomic coordinates and B factors) will have associated uncertainties. These arise not only from systematic and random errors in the experimental data but also in the interpretation of those data. Currently, the uncertainties cannot easily be estimated for macromolecular structures due to the computer- and memory-intensive nature of the calculations required (Tickle *et al.*, 1998). Thus, more indirect methods are necessary to assess the reliability of different parts of the model, as well as the reliability of the model as a whole. Among these methods are those which rely on checking only the stereochemical and geometrical properties of the model itself, without reference to the experimental data (MacArthur *et al.*, 1994; Laskowski *et al.*, 1998). Here we describe *PROCHECK* (Laskowski *et al.*, 1993), which is one of these structure-validation methods.

The *PROCHECK* program computes a number of stereochemical parameters for the given protein model and compares them with 'ideal' values obtained from a database of well refined high-resolution protein structures in the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The results of these checks are output in easy-to-understand coloured plots in PostScript format (Adobe Systems Inc., 1985). Significant deviations from the derived standards of normality are highlighted as being 'unusual'.

The program's primary use is during the refinement of a protein structure; the highlighted regions can direct the crystallographer to parts of the structure that may have problems and which may need attention. It should be noted that outliers may just be outliers; they are not necessarily errors. Unusual features may have a reasonable explanation, such as distortions due to ligand binding in the protein's active site. However, if there are many oddities throughout the model, this could signify that there is something wrong with it as a whole. Conversely, if a model has good stereochemistry, this alone is not proof that it is a good model of the protein structure.

references