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. 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.

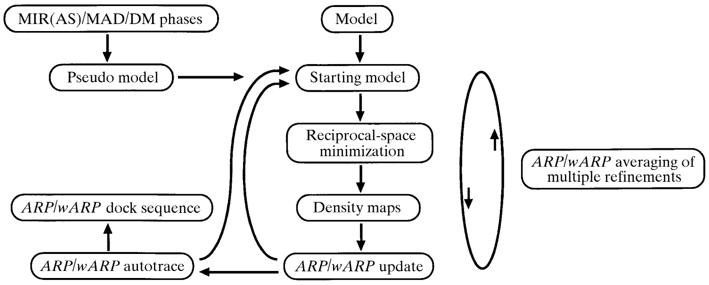

Fig. 25.2.5.1. A flow chart of the *Automated Refinement Procedure*.

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,