and/or strands of unknown connectivity and uncertain directionality) can be fed into another program, *DEJAVU* (Kleywegt & Jones, 1994*b*, 1997*b*) (see below), to check if they are similar to (a part of) another protein whose structure is known (Kleywegt & Jones, 1994*b*).

Finally, *RAVE* also contains three utility programs that can be used to manipulate three essential data structures encountered in averaging, map interpretation and refinement:

(1) *MAMA* (Kleywegt & Jones, 1994*b*, 1999). This program is used to generate, analyse and manipulate masks (molecular envelopes). It contains many of the tools described earlier by Jones (1992), but many new features have been added to it since. Masks can be generated from scratch, using a PDB file or *BONES* atoms, by 'recycling' another mask (*e.g.* from a different crystal form), or by combining several older masks. The quality of masks can be improved by filling voids, removing unconnected 'droplets', smoothing the surface, trimming regions that give rise to overlap through (non-)crystallographic symmetry and checking that all atoms in a model are covered by the mask (*e.g.* after changes or extensions to a model have been made).

(2) *MAPMAN* (Kleywegt & Jones, 1996*a*). This program was written for format conversion, analysis and manipulation of electron-density maps. Maps can be read and written in a variety of formats, including those used by *O*. Maps can be combined, scaled, peak-picked and subjected to 'digital image filters' (Kleywegt & Jones, 1997*a*). Several older stand-alone programs have been incorporated into *MAPMAN*, such as *MAPPAGE* and *BONES* (Jones & Thirup, 1986; the program previously used to skeletonize electron density for use with *Frodo* or *O*). Many statistics and types of histograms and plots (*e.g.* slices, or 2D and 1D projections) can be calculated or generated.

(3) *DATAMAN* (Kleywegt & Jones, 1996*a*). This program is used for simple format conversion, analysis and manipulation of reflection data sets [consisting of Miller indices, $F$, $\sigma(F)$, and possibly a cross-validation flag]. Data can be sorted, Laue symmetry can be applied, and data can be scaled by a temperature and a scale factor, re-indexed, and reduced in special cases where higher symmetry is present or suspected. The program contains a wide range of options to select 'test set' reflections that are to be set aside for cross-validation purposes (Brünger, 1992*a*; Kleywegt & Brünger, 1996). Many statistics and types of histograms and plots can be calculated or generated.

## 17.1.4. Structure analysis

A number of programs are available for the analysis of protein models. Most of these programs are interfaced with *O*, producing files (such as maps and *O* macros) that allow for quick and easy visualization and inspection of their results.

(1) *VOIDOO* (Kleywegt & Jones, 1994*a*). This program can be used to find cavities in macromolecular structures. The program will detect cavities, measure their volumes and produce files that can be used by *O* to visualize the cavities, any atoms inside them and protein residues surrounding them.

(2) *DEJAVU* (Kleywegt & Jones, 1994*b*, 1997*b*). This is a program for fold recognition. It uses an abstracted representation of protein structure, namely the coordinates of the start and end points of secondary-structure elements (SSEs). The user can define a motif of SSEs (*e.g.* a four-helix bundle that represents only one domain of a much larger molecule) and the program will search a database derived from the PDB (Bernstein *et al.*, 1977) to look for other protein structures that contain a similar structural motif. Alternatively, the program can take all SSEs of a structure into account and look for structures in the database that have as many SSEs as possible in an arrangement similar to the user's structure. This approach led to the discovery of the structural similarity between glutathione synthetase and D-alanine:D-alanine ligase (Fan *et al.*, 1994, 1995), and that between the N-terminal DNA-binding domain of the diphtheria toxin repressor and the C-terminal DNA-binding domain of catabolite gene activator protein (Qiu *et al.*, 1995). Although several other fold-recognition programs are available, *DEJAVU* has two finesses that distinguish it, namely, the fact that it does not require that atomic coordinates (*e.g.* of the C$\alpha$ atoms) be available, and the fact that the program can use SSEs even in cases where their directionality and/or connectivity is unknown. This situation typically occurs in early stages of the map interpretation process, when some SSEs can be discerned in the density [*e.g.* using *ESSENS* (Kleywegt & Jones, 1997*a*)], but their direction is still difficult to determine, and when their connections may still be uncertain. In that case, the start and end points (in either order) can simply be estimated by the crystallographer (or using the program *SOLEX*, see Section 17.1.3) and the set of SSEs can be used as input to *DEJAVU* to look for similar structures in the database (Kleywegt & Jones, 1994*b*). Although the search will be less sensitive in this case, if successful, it may shorten the initial model-building process considerably.

(3) *SPASM* (Kleywegt, 1999*a*). This program is related to *DEJAVU*, but operates at a more detailed level. It can be used to look for known structures that contain a user-defined motif, consisting of two or more protein residues. This program has been used, for instance, in the analysis of phosphoenolpyruvate carboxykinase (Matte *et al.*, 1996), where it revealed a striking similarity between this protein's P-loop structure and the P-loops of adenylate kinase isozyme III, RecA protein and p21ras. We have also developed a program (*RIGOR*) that takes the opposite approach: using a database of pre-defined motifs (*e.g.* ligand and metal-binding sites, catalytic centres), this program will check if any of these also occur in the user's protein model.

(4) *SBIN* (Kleywegt & Jones, 1998; Kleywegt, unpublished programs). *SBIN* is a suite of programs that can be used to derive *PROSITE* patterns (Bairoch & Bucher, 1994) or Gribskov-style sequence profiles (Gribskov *et al.*, 1987; Gribskov & Veretnik, 1996) from sets of superimposed protein models. These patterns and profiles in turn can be used to retrieve protein sequences from databases such as *SWISS-PROT* and *TrEMBL* (Bairoch & Apweiler, 1997) that may be related in structure and/or function.

## 17.1.5. Utilities

Many utility programs are available from Uppsala, most of them aimed at practising crystallographers. Some of these (*MAMA*, *MAPMAN*, *DATAMAN*) have been discussed in Section 17.1.3. A few others are discussed below.

(1) *LSQMAN* (Kleywegt & Jones, 1997*b*; Kleywegt, 1996). This is a program for analysing and manipulating multiple copies of a molecule or multiple molecules. It contains tools to superimpose molecules (including an option to find such superpositioning automatically), to improve the fit of two superimposed molecules in myriad ways, to calculate and plot r.m.s. distances and $\varphi, \psi$ or $\chi_1, \chi_2$ torsion-angle differences and circular variances (Allen & Johnson, 1991; Korn & Rose, 1994; Kleywegt, 1996) of different molecules, to generate multiple-model Ramachandran plots, to compare the solvent structure in two molecules, to find the 'central' molecule of an ensemble [defined as the molecule that has the smallest r.m.s.(r.m.s.d.), *i.e.* the r.m.s. value of its pairwise r.m.s.d.'s to each of the other molecules], and to align molecules in an ensemble to the 'central' molecule. The program can handle proteins, nucleic acids and other types of molecules.

(2) *MOLEMAN*2. This is a general program for analysis and manipulation of molecules (in PDB-format files). It contains too