Acta Cryst. (1998). D**54**, 1119–1131

Databases in Protein Crystallography

Gerard J. Kleywegt* and T. Alwyn Jones

Department of Molecular Biology, Uppsala University, Biomedical Centre, Box 590, SE-751 24 Uppsala, Sweden. E-mail: gerard@xray.bmc.uu.se

(*Received 27 February 1998; accepted 18 May 1998*)

Abstract

Applications of structural databases in the protein crystallographic structure determination process are reviewed, using mostly examples from work carried out by the authors. Four application areas are discussed: model building, model refinement, model validation and model analysis.

1. Abbreviations

3D, three-dimensional; CBH, cellobiohydrolase; CSD, Cambridge Structural Database; HIC-Up, Heterocompound Information Centre, Uppsala; PDB, Protein Data Bank; RBP, retinol-binding protein; RMSD, root-mean-square distance or deviation; VRML, virtual reality modelling language; WWW, World Wide Web.

2. Introduction

An unattributed inequality postulates that 'data $\neq$ information $\neq$ knowledge $\neq$ wisdom'. To proceed from raw data to information requires processing of that data (*e.g.*, calculation of the most densely populated areas of the Ramachandran plot based on a set of 1000 high-resolution protein models taken from the PDB). To translate information into knowledge requires careful analysis, interpretation, and validation (*e.g.*, the knowledge that the large majority of residues in protein structures lie in one of the most densely populated areas of the Ramachandran plot, with the exception of glycine residues). To proceed from knowledge to wisdom requires insight (*e.g.*, to apply Occam's razor when faced with a model with a poor Ramachandran plot, namely to assume that the model contains problematic regions, rather than assuming it to be the result of a freak of Nature, or even due to El Niño). Over the past decade, databases, and information and knowledge derived from databases, have become indispensable tools for practising macromolecular crystallographers.

Nowadays, at most stages of a structure determination project databases are used, either explicitly or implicitly (*e.g.*, using information or knowledge derived from analysis of databases). A typical project may start with a literature search using, for example, the MEDLINE or ISI databases. If the sequence of a target protein is available, a wealth of sequence comparison and analysis tools can be used. For example, to find (globally or locally) homologous proteins, programs such as *BLAST* (Altschul *et al.*, 1990) or *FASTA* (Pearson & Lipman, 1988) can be used on large protein and/or translated nucleic acid sequence databases, such as SWISS-PROT and TrEMBL (Bairoch & Apweiler, 1997), and GenBank (Denson *et al.*, 1997). To identify sequence characteristics associated with structure or function, the PROSITE (Bairoch & Bucher, 1994) or ProDOM (Sonnhammer & Kahn, 1994) databases can be accessed. When the time has come to produce diffraction-quality crystals, crystallization and heavy-atom databases can be consulted. If a protein is similar in sequence to another one whose structure is known, that structure can be retrieved from the Protein Data Bank (Bernstein *et al.*, 1977), and used as a probe in molecular-replacement calculations. In other cases, a model may have to be built from scratch using experimental electron density, a task typically involving the recycling of fragments found in a (small) structural database. When a model is sufficiently complete to be subjected to crystallographic refinement, target values for its geometry can be derived from an analysis of high-resolution small-molecule crystal structures as found in the CSD (Allen *et al.*, 1979). During the rebuilding and refinement process, database methods can be used to check the progress and to pinpoint parts of the model that may be problematic. Similar tools can be used to validate the final model, prior to deposition and publication (MacArthur *et al.*, 1994). In the final stage, while analysing the structure, databases can be used to look for similarities with other proteins whose structure is known, be it at the level of the overall fold (Holm & Sander, 1994; Kleywegt & Jones, 1997*c*), or at the level of, *e.g.*, loops and active-site residues (Kleywegt, 1998).

Here we review some of the methods and databases used in the actual process of protein model building, refinement, validation, and analysis. In addition, we briefly describe some of our recent work in these areas.

3. Model building

In the early days of protein crystallography, protein models were built by hand (Kendrew *et al.*, 1960), using

metal rods to show chemical bonds. Although myoglobin was built with a series of vertical metal rods that were colour-coded to represent density values, later wire models used an optical system based on an inclined semi-silvered mirror to produce an illusion of superimposing a contoured electron density onto the model (Editorial, 1997). The model was supported by a series of rods and clamps. With the advent of affordable computers and graphics systems, numerous software systems were designed to replace these wonders of engineering.

The first application of structural databases in the area of crystallographic model building was described by Jones & Thirup (1986). It was initially developed in order to make the generation of a trace from a skeleton simpler and more effective, Fig. 1. [A few years earlier, Jonathan Greer (1981) had used multiple protein models for the construction of homology models.] When the method [implemented in *FRODO* (Jones, 1978, 1985)] was tested, it was noticed that two turns in retinol-binding protein (RBP) were very similar in structure, yet did not resemble any previously classified type of turn. A search of the PDB revealed many more instances of this type of turn, which triggered the question whether any part of the RBP structure was unique, or whether the whole RBP structure could be constructed through recycling of fragments from other, previously solved protein structures. Indeed, as it turned out, RBP could be reconstructed from only three other protein structures with an RMSD of the order of 1 Å on Cα atoms. The next step, then, was to create a database consisting of a small number (initially 37) of well refined high-resolution protein structures that could be used to construct new protein models using crystallographic data (Jones & Thirup, 1986), NMR data (Kraulis & Jones, 1987), or in homology modelling (Jones & Thirup, 1986).

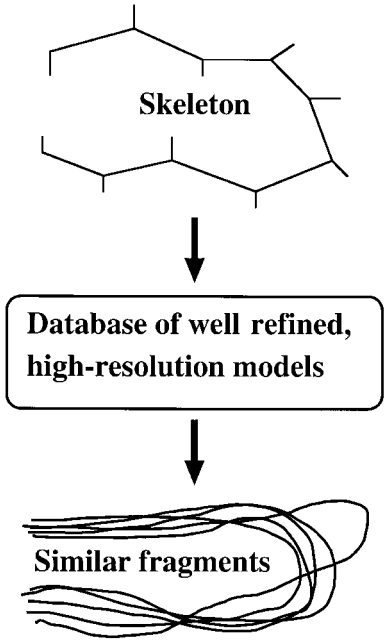


Fig. 1. Illustration of the use of structural databases to generate main-chain coordinates for a protein model, based on skeletonized electron density.

Before automatic model-building procedures can be used, a set of 'guide points' is required. In the case of electron-density maps these are conveniently abstracted in the form of a so-called skeleton (Greer, 1974, 1985). In the original implementation in *FRODO*, such a skeleton had to be converted into a set of Cα positions, either automatically (by placing points along the skeleton at 3.8 Å intervals), or manually. This initial set of Cα positions could then be used to query the structural database. For reasons of speed, least-squares superpositioning methods were impractical at the time, and so a two-step procedure was used. In the first step, a simple method based on C$\alpha-$Cα distance plots (Phillips, 1970) (*i.e.*, matrices containing the distances between all pairs of Cα atoms in a protein model) was used to locate fragments that were likely to be similar. In the second step, a full least-squares analysis was used on the selected fragments. The distance matrices were precomputed for all structures in the database, and locating fragments of similar local conformation to that of a stretch of N guide Cα positions was, therefore, a simple and speedy operation. For each consecutive stretch of N residues in the database structures, the sum of squared differences between the inter-Cα distances was calculated. The database fragments for which this sum was small were then used in the least-squares comparison. Originally, the length of the fragments could be determined by the user (this method is still available in *O* as the *Lego_CA* command). Later, this was fixed at five residues, which turned out to be sufficient to reproduce main-chain coordinates with an RMSD of ~0.5 Å (Jones *et al.*, 1991). This cut-off, in turn, ensures that the carbonyl O atoms will be pointing in the right direction in most instances. An additional benefit of using shorter fragments is that less-frequent main-chain conformations have a higher probability of being recognised.

The current implementation [the *Lego_auto_mc* command in *O* (Jones *et al.*, 1991; Jones & Kjeldgaard, 1994, 1997)] locates the best fit for five-residue stretches in the database ($i-2$ to $i+2$), but it only updates the coordinates of the middle three residues ($i-1$ to $i+1$). The algorithm then moves forward three residues and finds the best fit for residues $i+1$ to $i+5$, *etc.* In this fashion, it rapidly generates a set of main-chain coordinates for a model, starting from approximate Cα positions. [If the random error in the approximate Cα positions is greater than ~0.3 Å, the autobuilt model will be closer to the true structure than the starting model (Jones *et al.*, 1991).] A side-effect of the use of five-residue fragments to generate coordinates for three residues at a time is that all residues other than number

3, 6, 9, *etc.* will have their main-chain φ and ψ torsion angles determined by the fusion of two fragments that are not necessarily adjacent fragments from one and the same database structure, Fig. 2. Hence, paradoxically, models generated in this fashion (*i.e.*, derived entirely from recycled database fragments) will generally not display a Ramachandran plot typical of a well refined high-resolution model, even though all the structures in the database had good Ramachandran plots. However, because the random errors in the main chain are then usually rather small, a single cycle of crystallographic refinement quickly leads to a much improved Ramachandran plot (Kleywegt & Jones, 1996*b*), Fig. 3.

The algorithm outlined here lies at the basis of many a homology modelling program. Interestingly, the approach was also extended for application to NMR data (short and medium range NOEs plus vicinal coupling constants) (Kraulis & Jones, 1987). In this case, a slightly larger database was used (56 protein crystal structures refined to a resolution of 2.0 Å or better), and instead of using Cα—Cα distance matrices, distances between calculated H*N*, Hα and Hβ protons were used. Not unexpectedly, the approach produces models with good local conformations, but since the long-range NOE information is not used, the relative orientation of secondary-structure elements, for instance, is ill-determined. Nevertheless, the approach showed promise as a method for local refinement of structures generated by other means (*e.g.*, distance-geometry or simulated-annealing methods). This method has not caught on in the NMR community, but other methods have been developed to make NMR models more 'protein-like' (see below).

4. Side chains

About two decades ago, Joel Janin and co-workers investigated the distribution of χ_1 and χ_2 side-chain torsion angles in crystal structures of proteins (Janin *et al.*, 1978). They found that these torsion angles behave in accordance with simple energy-based calculations, with preferences for values of +60, 180 and −60° for aliphatic side chains, and +90° and −90° for the χ_2 torsion angle of aromatic residues. Inspection of the combined χ_1/χ_2 distributions revealed that several types of residue displayed preferences for certain combinations of torsion angles. For example, leucine residues turned out to prefer the combinations −60/180° and 180/

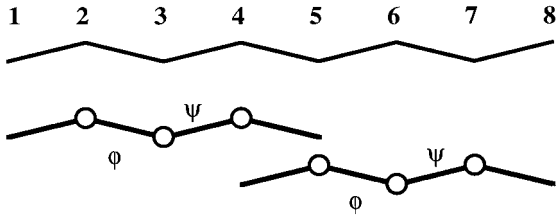


Fig. 2. When *O* auto-builds main-chain coordinates, overlapping fragments of five residues are retrieved from the database, and these are used to update the coordinates for the central three residues. Hence, only every third residue will inherit main-chain torsion angles (φ and ψ) from a single database fragment (see also Fig. 3).

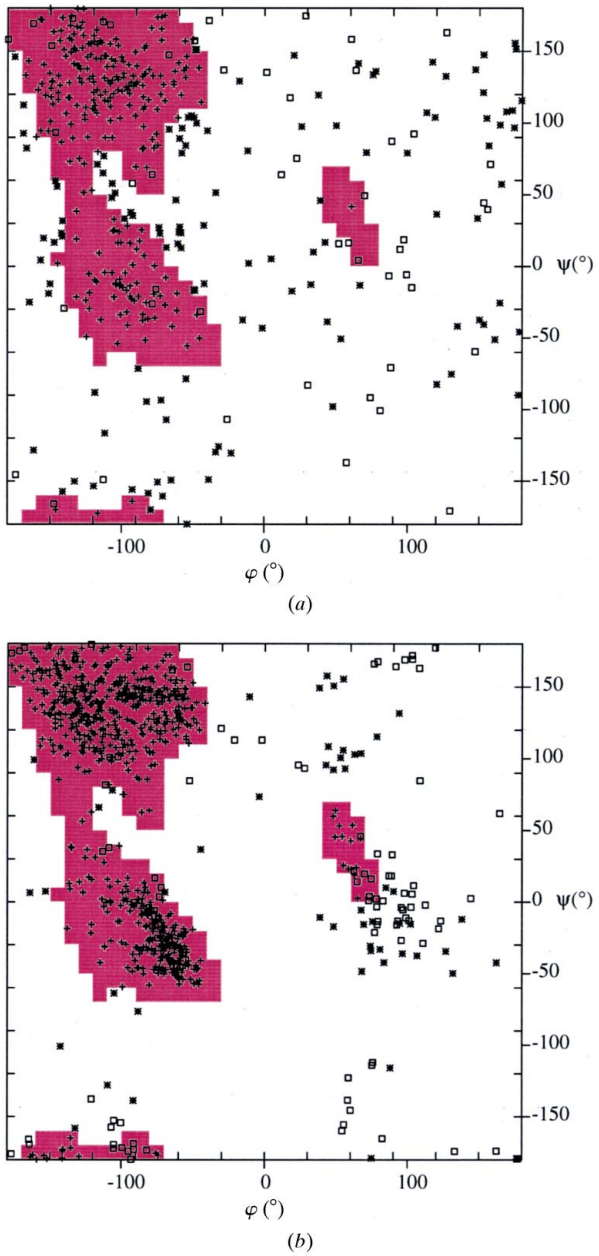


Fig. 3. The phenomenon illustrated in Fig. 2 explains why the Ramachandran plot of an automatically built model is usually poor. However, the errors tend to be small and randomly distributed, and therefore the model can usually be vastly improved by a single cycle of reciprocal space crystallographic refinement. Ramachandran plot of (*a*) the initial model of cellobiohydrolase I (Divne *et al.*, 1994), and (*b*) the model obtained after a single cycle of simulated-annealing refinement. In both plots, the pink areas represent the core regions as defined in (Kleywegt & Jones, 1996*b*).

Table 1. *Distribution of individual side-chain torsion angles*

We used the list of Hobohm & Sander (1994) of August 1995 and the Protein Data Bank release of October 1995, to create a set of 403 protein models that were mutually 95% or less identical in sequence, that contained more than 20 amino-acid residues, and that had been solved by X-ray crystallography at a resolution not worse than 2.0 Å. For each model, all atoms (and their associated torsion angles) whose temperature factor was higher than the average protein temperature factor plus two standard deviations were discarded. This was performed in order to exclude residues from the analysis whose conformation might have been determined more by the restraints or force field used in the refinement than by actual experimental data. For the side-chain torsion-angle analysis, each torsion angle was divided into bins of 1°. The preferred torsion-angle values were integrated and averaged with an interactive graphics program, *O2D* (GJK, unpublished program). In the table, the average values of the most populated regions (that account for at least 5% of the population) are listed in degrees, together with the percentage of the population they represent (in parentheses).

Residue	Torsion	Preferred values and population (%)		
Thr	χ_1	300 (46)	61 (43)	187 (7)
Cys	χ_1	295 (55)	182 (28)	63 (16)
Ser	χ_1	63 (45)	297 (30)	180 (21)
Val	χ_1	175 (69)	297 (19)	66 (7)
Leu	χ_1	289 (65)	181 (29)	
	χ_2	175 (58)	64 (30)	
Pro	χ_1	267 (45)	337 (43)	
	χ_2	326 (46)	34 (44)	
His	χ_1	294 (55)	185 (32)	61 (12)
	χ_2	278 (47)	82 (34)	178 (7)
Ile	χ_1	296 (74)	61 (14)	191 (10)
	χ_2	168 (78)	299 (14)	78 (7)
Phe	χ_1	293 (52)	181 (33)	62 (13)
	χ_2	77 (45)	282 (40)	
Tyr	χ_1	293 (53)	180 (32)	63 (13)
	χ_2	77 (48)	282 (43)	
Trp	χ_1	291 (53)	181 (32)	61 (14)
	χ_2	91 (57)	263 (27)	342 (8)
Asp	χ_1	290 (50)	187 (30)	61 (18)
	χ_2	one broad peak	(max. ~340°)	
Asn	χ_1	291 (52)	188 (29)	63 (15)
	χ_2	flat distribution		
Met	χ_1	292 (60)	185 (28)	63 (7)
	χ_2	179 (60)	296 (25)	68 (10)
	χ_3	286 (40)	74 (31)	180 (19)
Glu	χ_1	292 (55)	185 (30)	62 (9)
	χ_2	179 (63)	290 (19)	73 (14)
	χ_3	one broad peak	(max. ~0°)	
Gln	χ_1	294 (57)	183 (31)	66 (6)
	χ_2	178 (61)	292 (20)	69 (14)
	χ_3	flat distribution		
Arg	χ_1	292 (55)	183 (31)	66 (8)
	χ_2	181 (75)	289 (11)	67 (6)
	χ_3	179 (43)	293 (27)	68 (21)
	χ_4	176 (46)	266 (28)	93 (21)
Lys	χ_1	291 (56)	185 (32)	64 (6)
	χ_2	179 (71)	289 (14)	69 (7)
	χ_3	178 (70)	292 (11)	69 (9)
	χ_4	177 (59)	292 (17)	68 (15)

+60°. Upon completing their high-resolution refinement of penicillopepsin, James & Sielecki (1983) repeated the analysis using a small set of well refined high-resolution crystal structures from their own laboratory. This revealed the distributions of torsion angles and combinations to be even narrower and sharper than had been observed previously.

A few years later, Ponder & Richards (1987) derived a library of (preferred) side-chain rotamers, *i.e.* residue-specific preferred (combinations of) side-chain torsion angles, for the purpose of enumerating sequences that could effectively pack on a given backbone scaffold or 'core structure'. This set of rotamers formed the basis for the rotamer library used in *O* (Jones *et al.*, 1991), which retained only those rotamers that occurred with a frequency of at least 10% in the analysis of Ponder and Richards, and which mostly used the χ_1 and χ_1/χ_2 torsion angles. When *O* autobuilds side chains, every residue is modelled by default in its most common rotamer conformation (the *Lego_auto_sc* command). Subsequently, the user can correct those instances where the side chain is in a different rotamer conformation (*Lego_side_ch* command) or in a non-rotamer conformation (*Tor_residue* and *Tor_general* commands). In the former case, the program can also execute this task automatically (*RSR_rotamer* command) by calculating for each rotamer how well it fits the experimental electron density (after a rigid-body rotational search pivoting around the Cα atom in order to optimize the fit to the density). The rotamer conformation that gives the best fit is subsequently selected. More recently, a new command has been added to *O* that also allows automatic real-space fitting of torsion angles against the density [*Fm_rsr_tors* command, (TAJ, unpublished results)].

We have recently repeated the analysis of side-chain torsion angles, now using a 5% population cut-off to obtain a larger set of rotamers (Tables 1 and 2). One interesting observation pertains to the third leucine rotamer, which has rather unusual torsion angles, yet accounts for almost 10% of the leucine population surveyed, Fig. 4. It is most likely that this is a (frequent) model-building artifact, since its shape resembles that of the most frequent rotamer. This pitfall has also been noted by P. A. Karplus (quoted and discussed in Kuszewski *et al.*, 1997). This case may serve as a warning for crystallographers, homology modellers, and structure validators.

Other workers have derived rotamer libraries that take into account a dependence on the local main-chain conformation. However, for crystallographic model-building purposes this is unnecessary, since the correct conformation can usually be identified on the basis of the shape of the electron density (*caveat emptor*; see above).

5. Bias

We have discussed two ways of using databases in the process of protein crystallographic model building: the construction of a model's main chain using a small structural database, and the generation of side chains

Table 2. *Rotamer library used for crystallographic model building with O*

The data described in the headnote to Table 1 were used to generate χ_1, χ_2 plots, that were divided into $10 \times 10^\circ$ squares, and the torsion-angle combinations were tallied for each residue type individually. Rotamer populations and values were obtained by integration and averaging with *O2D*. Preferred torsion angles are listed as g^- ($\sim+60^\circ$), g^+ ($\sim-60^\circ$), t ($\sim180^\circ$), c ($\sim0^\circ$), o^+ ($\sim+90^\circ$), o^- ($\sim-90^\circ$) and x (other values); $o^\pm$ is used in cases where a χ_2 torsion angle of $+90^\circ$ is chemically equivalent to one of -90°. Only rotamers whose population is at least 5% have been included. Glycine and alanine do not have side-chain rotamers. Preferred conformations for cysteine, valine, threonine, and serine side chains are defined by the χ_1 torsion alone, and are listed in Table 1.

Residue	Rotamer	Population (%)	χ_1 (°)	χ_2 (°)	Residue	Rotamer	Population (%)	χ_1 (°)	χ_2 (°)
Leu	g^+t	54	300	180	Asn	g^+x	33	300	320
	tg^-	27	190	70		tc	21	190	0
	o^-x	9	260	40		g^+x	17	300	140
Pro	xx	40	30	303		g^-c	12	70	0
	xx	37	340	40		tx	7	190	210
	cc	20	0	0	Met	g^+t	36	300	190
His	g^+g^+	30	300	290		g^+g^+	26	300	300
	g^+o^+	19	300	100		tt	17	190	180
	to^+	16	190	80		tg^-	9	190	70
	to^-	12	190	270		g^-t	7	70	190
	g^+t	8	300	170	Glu	g^+t	34	300	180
	g^-o^-	7	70	280		tt	23	190	180
Ile	g^+t	58	300	170		g^+g^+	15	300	300
	tt	7	190	170		g^-t	6	70	190
Phe	$g^+o^\pm$	49	300	90	Gln	g^+t	38	300	180
	$to^\pm$	32	190	90		tt	20	190	180
	$g^-o^\pm$	13	70	90		g^+g^+	18	300	300
Tyr	$g^+o^\pm$	48	300	90		tg^-	10	190	70
	$to^\pm$	32	190	90		g^-t	5	70	100
	$g^-o^\pm$	14	70	90	Arg	g^+t	44	300	190
Trp	g^+o^+	38	300	100		tt	27	190	190
	to^+	15	180	80		g^+g^+	11	300	290
	to^-	14	180	260		g^-t	8	70	190
	g^+c	9	300	350		tg^-	5	190	70
	g^-o^-	9	70	280	Lys	g^+t	42	300	190
	g^-o^+	5	70	90		tt	25	190	180
	g^+o^-	5	300	280		g^+g^+	14	300	290
Asp	g^+x	39	300	340		g^-t	6	70	180
	tx	14	190	340		tg^-	6	190	70
	g^-x	11	70	340					

using a derived database (rotamer library). It is sometimes argued that the use of databases in model building will introduce bias into the model (and, hence, the structural database), making it more likely that genuinely unusual conformations will be discriminated against. We submit that this argument is invalid. First, in regions where the experimental density is of high quality, the crystallographer (or a refinement program) can and will rely on this density to model the protein's main chain and side chains. In regions where the density is ambiguous or even invisible, one has no proof for the model assuming a 'genuinely unusual' conformation in the first place. In this case, it is safer to rely on a database of well refined and high-resolution structures to model the main chain, and on a library of rotamers to model the side chains. Frequently, unusual local conformations arise because of model-building errors (see Kleywegt *et al.*, 1996, for an example). Also, as more and more structures are solved at atomic resolution and refined without reference to databases it becomes clear that protein structures are even more 'well behaved' (in terms of main-chain and side-chain geometry) than assumed previously (see, *e.g.*, Sevcik *et al.*, 1996). The use of databases helps the crystallographer to rapidly produce a 'zero-order' model that can be expected to be reasonably close to the final model (Jones & Thirup, 1986; Jones *et al.*, 1991; Zou & Mowbray, 1994), and that portrays genuine protein-like features. [Or, as Kuszewski and co-workers have argued (Kuszewski *et al.*, 1996), the distributions of main-chain and side-chain torsion angles found in the crystallographic database are a direct result of the underlying physical chemistry of the system.] Careful refinement and intelligent rebuilding (Kleywegt & Jones, 1995*a*, 1996*a*, 1997*b*) will subsequently apply the final touches, as well as reveal parts of the protein that genuinely deviate from the zero-order model. It should be noted that such deviations are often of biological importance (Herzberg & Moult, 1991). For instance, the active-site serine residue in α/β hydrolase enzymes (Ollis *et al.*, 1992) is an outlier in the Ramachandran plot, but this is a feature of these proteins, not an error in the models. In order to be able to discriminate errors and genuine outliers, one often needs access to the experimental data (Jones *et al.*,

1996). Deposition of structure factors is therefore crucial to ensure the integrity of the structural database.

6. Refinement

In crystallographic refinement, database information is typically used in the form of dictionaries that describe geometrical and stereochemical features (*e.g.*, bond lengths and angles, planarity and chirality) in terms of target values and the (desired or observed) tightness of the distribution around these target values (Hendrickson & Konnert, 1980). This information is used to restrain or constrain the model during crystallographic refinement, since the crystallographic data alone (at anything worse than atomic resolution) contain insufficient information to produce proper geometry. If one were to refine a model only against the crystallographic data, this would lead to spurious 'errors' and a general deterioration of the model, as indeed it did in the very first published least-squares refinement of a protein model, that of rubredoxin at 1.5 Å resolution (Watenpaugh *et al.*, 1973). Until a few years ago, all refinement and model-building programs used their own set of atom types and geometric target values (Laskowski, Moss *et al.*, 1993; Priestle, 1994). Engh & Huber (1991) carried out an analysis of the geometry of fragments of small molecules, as found in the CSD, that resemble fragments occurring in the 20 amino acids. This yielded a new set of unique atom types, new target values for bond lengths and bond angles and values for their experimentally observed sample standard deviations. This prompted Brünger to use the free *R* value (Brünger, 1992; Kleywegt & Brünger, 1996) to determine appropriate weights for the geometric terms relative to the weight of the crystallographic residual (Brünger, 1993). The improved target values for bond lengths and bond angles, combined with the fact that they were restrained more tightly, led to a general improvement of the quality of refined protein models with respect to both the bond lengths and the bond angles, and the fit to the crystallographic data as assessed using the free *R* value. Priestle has 'translated' the Engh & Huber dictionary for most of the common refinement and rebuilding programs (Priestle, 1994), and nowadays most protein models are refined with it. More recently, the group of Berman has carried out a similar analysis for nucleic acids yielding a much improved dictionary (Parkinson *et al.*, 1996), compatible with the Engh & Huber dictionary for proteins (*i.e.*, target values derived from a similar source, and restraints of comparable strength).

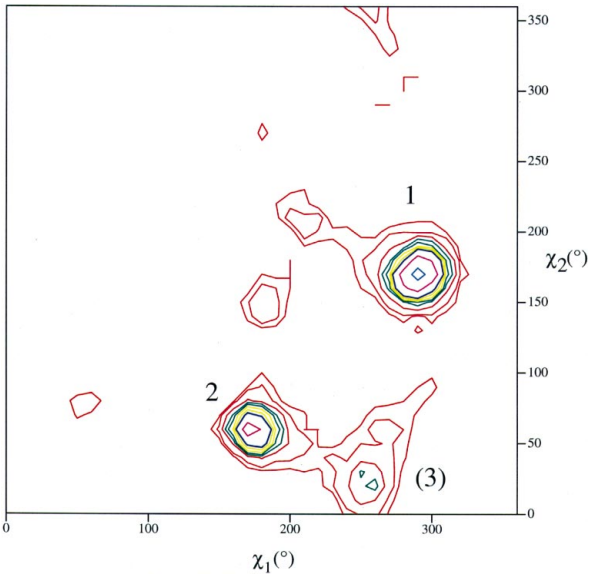


Fig. 4. Distribution of side-chain torsion angles for 6638 leucine residues as observed in 403 crystal structures. The two major rotamers are labeled 1 and 2, and the probably spurious third rotamer is labeled (3).

Compared with the high-quality dictionaries available for protein and nucleic acid model refinement, the dictionaries used for other entities ('hetero compounds') are generally in a sorry state (GJK, unpublished observations). Because of the unlimited chemical diversity of hetero compounds, compared with the small number of building blocks that make up proteins and nucleic acids, a comprehensive analysis in the vein of Engh & Huber is impractical. Every time a new hetero compound is introduced into a refinement or model-building program, dictionaries will have to be defined. Sometimes these can be derived from the entries for regular amino acids or nucleic acids, or they may be obtainable from colleagues (in which case they should be critically checked), or experienced chemists may be able to define them largely from scratch. Alternatively, the CSD can be searched to find out if the crystal structure of the hetero compound (or a related compound) has been solved previously. If this is not the case, the CSD can still be used to retrieve instances of smaller fragments of the hetero compound, and statistics pertaining to the distributions of bond lengths and angles can be calculated to yield target values and approximate standard deviation values. However, the CSD is a commercial database, and relatively few macromolecular crystallographers have access to it (although the Cambridge Crystallographic Data Centre operates a scheme under which infrequent academic users may be granted some access time free of charge). The PDB is an alternative database to look for coordinates of hetero compounds, and we have recently set up a WWW-based service for this purpose, called HIC-Up ('Hetero-compound Information Centre, Uppsala', at URL http://alpha2.bmc.uu.se/hicup/). This site contains coordinates, ready-made dictionaries (for *CNS*, *X-PLOR*, *TNT* and *O*), as well as other relevant information for the hetero compounds encountered in the PDB. The user should be aware, however, that macromolecular crystallography is

generally not a reliable method to determine small-molecule structures. Not only will limited resolution lead to less accurate hetero-compound structures, they may also have been refined using inappropriate dictionaries in the first place. In order to try and prevent indiscriminate use of dictionaries derived from such coordinates, a simple quality assessment is included for every hetero compound (see below).

A novel application of the use of databases in refinement is the approach of Kuszewski and co-workers, who have developed a database-derived conformational potential (Kuszewski *et al.*, 1996, 1997). In their 1996 paper, they noted that 'in most cases, a high-resolution ($\leq$ 2 Å) crystal structure will provide a better description of the structure in solution than the corresponding NMR structure' (for example, chemical shifts calculated from crystallographic models compare better with those determined experimentally than chemical shifts calculated from NMR models). This prompted them to devise a mechanism through which conformational information derived from high-resolution crystal structures can be incorporated into the (NMR) model-refinement process. Their original implementation used the *PROCHECK* database of high-resolution crystal structures (Laskowski, MacArthur *et al.*, 1993) to derive matrices of energy values at evenly spaced points along axes that correspond to the various types of dihedral angle found in proteins (*e.g.*, χ_1 angles, φ/ψ and χ_1/χ_2). The populations were counted in bins, converted into probabilities, and transformed into a pseudo-potential by taking the negative logarithm (derivatives are approximated simply by the local slope of the energy function). Models refined using this potential fit the NMR data equally well, but in addition they converge more rapidly and (not unexpectedly) they score much better in quality tests using *PROCHECK* and *WHAT IF*. Naturally, since this method essentially 'fudges' both the Ramachandran plot and the rotamer distributions, these two criteria can no longer be used to validate a model refined in this fashion!

The conformational database potential method has been implemented in the refinement program *CNS* (Brünger *et al.*, 1998). We have carried out some preliminary tests of the use of this potential in the refinement of a low-resolution protein model [endoglucanase I (Kleywegt *et al.*, 1997), at 3.6 Å resolution]. Using validation tests that are largely orthogonal to the potentials used in the refinement program (such as the free *R* value, pep-flip score, and Cα backbone quality), we find that the method has a modest but distinct effect when used on the final model. However, when used in the refinement of an early, incomplete and partially mistraced model, the effect is mostly cosmetic (*i.e.*, improved Ramachandran plot and rotamer quality, but no impact on independent quality measures), and might even lead to a false impression concerning the quality of the model. A plausible explanation for these observations is that the database potential forces a model to assume *favourable* torsion-angle combinations, but not necessarily the *correct* ones. In crude models, this will

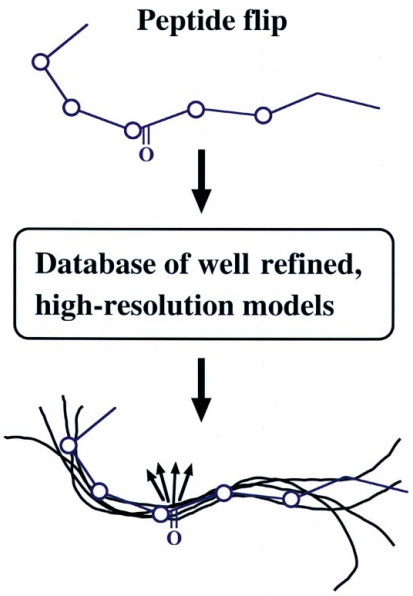


Fig. 5. Calculation of the pep-flip value as implemented in *O*.

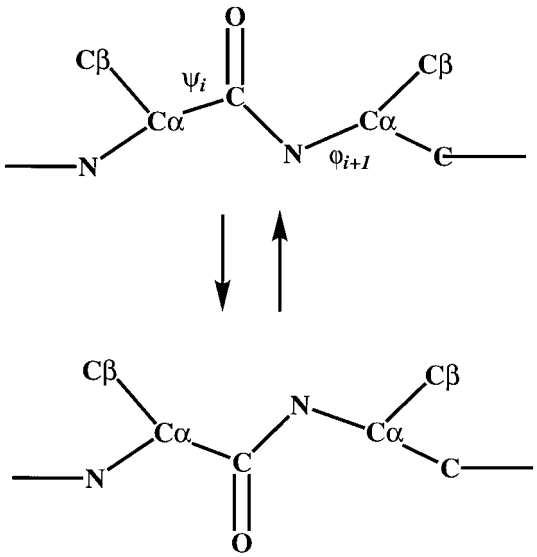


Fig. 6. The orientation of a peptide plane is intimately linked to the location in the Ramachandran plot of the two residues that are linked by it. Flipping the peptide plane between residues *i* and *i*+1 changes the ψ angle of residue *i* and the φ angle of residue *i*+1 by ~150–180°. If residue *i* has a negative φ value, an erroneous flip may not result in the residue becoming an outlier. However, if residue *i*+1 has a negative φ value, an erroneous flip will almost always result in the residue becoming an outlier. Hence, residues that have unusual φ, ψ values and are pep-flip outliers are often indicative of local main-chain errors.

lead to many residues lying in favourable but incorrect regions of the Ramachandran plot, for instance. On the other hand, on near-final models, in which most of the atoms are roughly in their correct position already, the effect is more benevolent. When deciding whether or not to use the database potential in crystallographic refinement, one will need to weigh the importance of improved model quality and the inevitable loss of several powerful validation criteria.

7. Validation

Structural databases have become an indispensable tool in the area of model validation. Ramachandran analysis is a prime example of this; whereas the 'allowed' areas of the Ramachandran plot were originally defined based on simulations of dipeptides (Ramachandran *et al.*, 1963; Ramakrishnan & Ramachandran, 1965), nowadays most programs use distributions derived from an analysis of a set of well refined and high-resolution crystal structures (Morris *et al.*, 1992; Laskowski, MacArthur *et al.*, 1993; Kleywegt & Jones, 1996*b*; Hooft *et al.*, 1997).

Recently, we developed a Ramachandran-like procedure for the validation of protein models for which only Cα coordinates are available (Kleywegt, 1997). It is based on the use of pseudo-angles and pseudo-torsion angles between sequential Cα atoms (Oldfield & Hubbard, 1994). A set of high-resolution models from the PDB was used to delineate core, 'disallowed', and other regions. It was shown that the fraction of residues in core and 'disallowed' regions are sensitive indicators of global model correctness, similar to the Ramachandran plot for all-atom models (Kleywegt & Jones, 1996*b*).

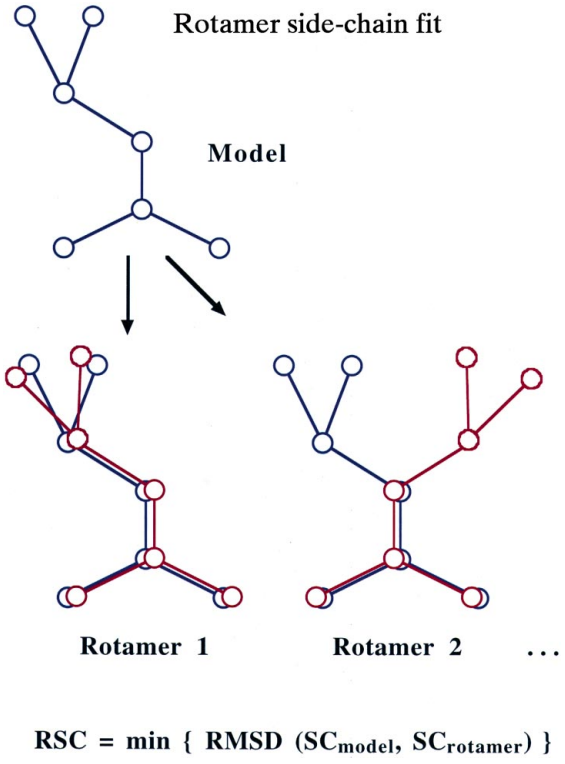


Fig. 7. Calculation of rotamer side-chain fit values as implemented in *O*.

In *O*, the same databases that are used in model building (see above) can be used to find local structural outliers, which may be either genuine, but unusual features, or errors. The quality of the main chain can be assessed quickly and sensitively by means of a Ramachandran plot (Kleywegt & Jones, 1996*b*). In addition, the orientation of the peptide O atoms can be investigated (*Pep_flip* command, Fig. 5). For every residue *i* in a model (except the two residues at each terminus), a penta-peptide *i*−2 to *i*+2 is used, and the structure database is searched to find up to 20 similar penta-peptides that superimpose with an RMSD of less than 1.0 Å on Cα atoms. The RMS distance of the peptide O atom of residue *i* to those of each of the database fragments is calculated and this number is called the pep-flip value. If the pep-flip value is large, the residue is classified as an outlier ('how large' depends on the size of the structural database; in *O*, typically, a value of 2.5 Å is used, but for a larger database a lower cut-off value would have to be used). This means that most residues in the database that have similar local Cα conformations have their carbonyl O atom pointing in the opposite direction to that of the model. This implies that the peptide plane has an unusual orientation, and it is up to the crystallographer to decide (using the electron density and/or analogy to related structures) if this is due to an error in the model, or whether it is a genuine feature of it. It is important to realise that almost every model contains a few outliers [typically, ~1–2% of the residues (Kleywegt & Jones, 1995*b*)]. As discussed in (Kleywegt, 1996), the orientation of the peptide plane is intimately associated with the location in the Ramachandran plot of the two residues linked by the peptide, Fig. 6.

The rotamer library of *O* can be used to pinpoint residues that have an unusual side-chain conformation (*RSC_fit* command, Fig. 7). For every residue (except glycyl and alanyl residues), each of the possible rotamers is superimposed using the main-chain coordinates, and the RMSD between the β, γ and δ side-chain heavy atoms is calculated. The RSC-fit value is defined as the RMSD of the rotamer that gives the smallest RMSD. If this number is large, the residue is classified as an outlier (again, 'how large' depends on the size of the rotamer library; for the original library a value of 1.5 Å was used, but for the enlarged library a value of 1.0 Å may be more appropriate). This implies that the residue does not have a side-chain conformation that resembles that of any preferred rotamer. Again, it is up to the crystallographer to investigate if this is a genuine feature of the

model, or due to an error. A typical final model will contain ~5–10% residues whose side chain is not in any preferred rotamer conformation (Kleywegt & Jones, 1995*b*). In the near future, the definition of rotamers in *O* will be recast in terms of the actual torsion angles, so that the RSC-fit value can be expressed as the RMS deviation of one, two or more torsion angles from preferred values or value combinations, as suggested by Noble *et al.* (1993).

As mentioned earlier, the geometry of hetero compounds in deposited models in the PDB is of widely varying quality. Since geometric dictionaries for such compounds often have to be formulated by the crystallographer, errors are easy to make and they will leave their mark on the final geometry of the ligand, co-factor, *etc.* Common mistakes include incorrect target values for bond lengths and bond angles, and omission of planarity and chirality restraints. Although this phenomenon has been observed previously (van Aalten *et al.*, 1996), few validation methods are available that are applicable to hetero compounds, and those that do exist tend to require access to the experimental data [*e.g.*, real-space electron-density fits (Jones *et al.*, 1991)]. In an attempt to provide at least a basic validation service, we have written a program (called *HETZE*) that checks whether bond lengths fall in a range of acceptable values [mainly using the information compiled by Allen *et al.* (1987), which is derived from an analysis of the CSD], whether torsion angles that would appear to be near 0 or 180° have been restrained sufficiently, and whether improper torsion angles (*i.e.*, virtual or pseudo-torsions, used by *X-PLOR* to enforce flatness and proper chirality) of C atoms with at least three non-H-atom neighbours assume reasonable values (near 0, +35 or −35°). This program is accessible through the HIC-Up WWW site mentioned earlier. The program has also been run on all the hetero compounds collected at that site to warn users for potentially unreliable coordinate sets.

Databases are used extensively in validation methods. The most interesting applications are those in which the criteria that are checked are orthogonal to the information included in the refinement (and rebuilding) process. One example of this is the 'directional atomic contact analysis' (DACA) method of Vriend & Sander (1993). This method, in essence, checks how usual or unusual the environment of each residue fragment is compared with the database. If there are a few residues with unusual environments in a model, this may help pinpointing interesting parts of the model. On the other hand, if many or most residues have unusual environments, then this is a strong indication that there is something seriously wrong with the model (*e.g.*, register error, tracing error, homology model). Other examples are methods that use pseudo-potentials or sequence–structure profiles to assess how likely the fold of the model is given the amino-acid sequence.

A large number of validation-related statistics have been collected for a subset of 476 crystallographic protein models from the PDB (Kleywegt, 1996). Although this Quality Data Base (QDB) was generated for the specific purpose of investigating the use of non-crystallographic symmetry in protein model refinement, it also provides information about many other validation criteria. A stand-alone program can be used to query the database, to sort entries by any criterion, and to investigate possible correlations between criteria [*e.g.*, between deviations from non-crystallographic symmetry and resolution (Kleywegt, 1996)]. It has also been used for formulating rules of thumb with respect to the expected percentage of outliers for several validation criteria (Kleywegt & Jones, 1995*b*).

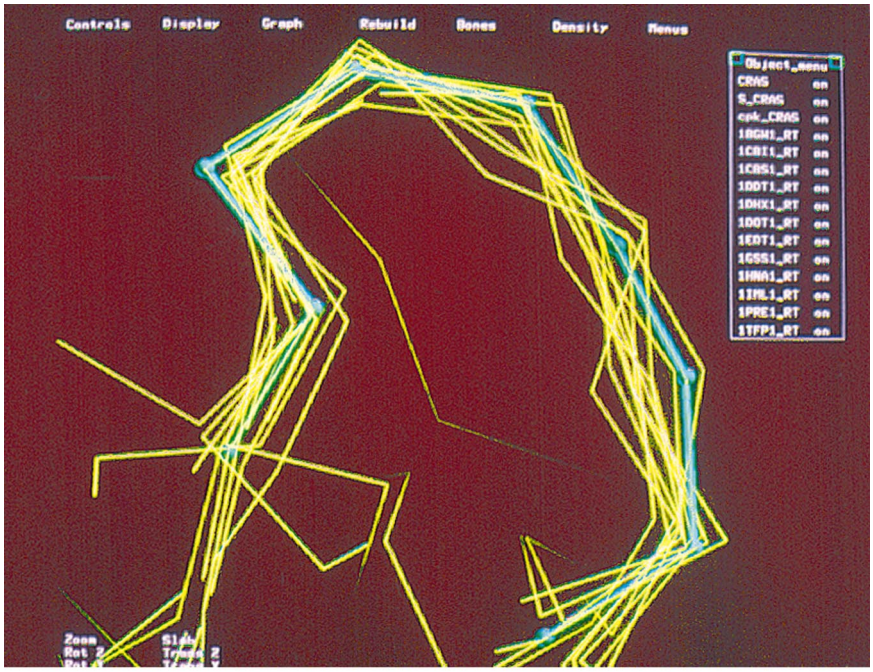
In the past, most validation tools have been based on the scrutiny of coordinates. For many of these tools, this means that only outliers can be identified (Jones *et al.*, 1996). In order to determine whether an outlier is due to a genuine but unusual feature of the molecule(s) under study, or whether it is more likely to be an error in the model, one often needs access to the original crystallographic data. This enables one to inspect maps, to calculate omit maps and, if necessary, to do some more refinement, perhaps using better methodology than was available when the model was originally refined. In Uppsala, Dr Tom Taylor is currently working on a project to link PDB entries to the crystallographic data (if deposited), as part of a European Union project on macromolecular model validation. This involves calculating maps that can be accessed through the WWW using VRML technology. At a later stage, limited refinement will also be carried out in order to obtain free *R* values and minimally biased maps.

8. Analysis

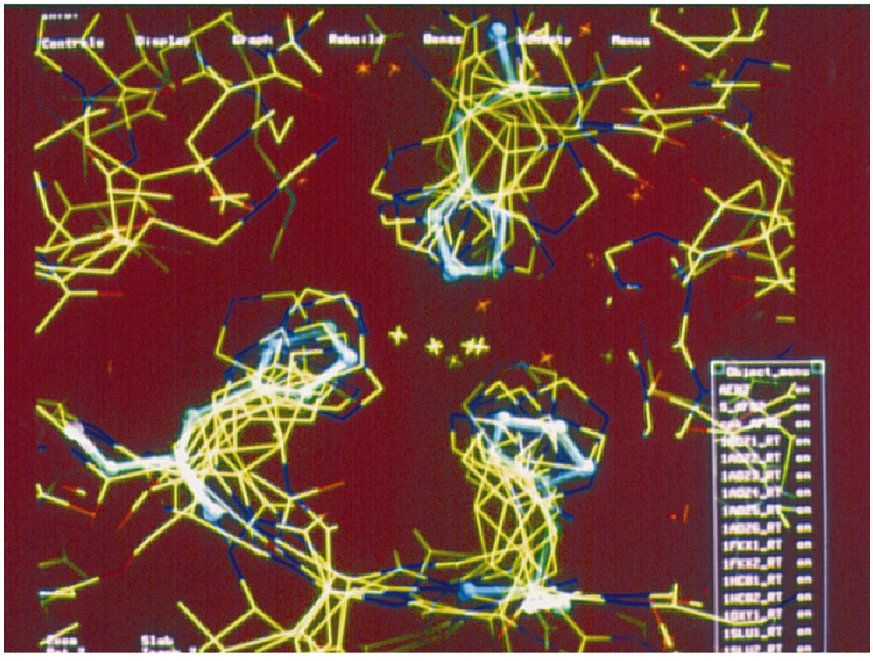
Once a protein structure has been solved, refined and validated, the hardest part of the job has been performed, but the biologically interesting work only begins. The first eagerly asked question will often be: does my protein have a novel fold? This question can be answered using any of a number of programs [such as *DALI* (Holm & Sander, 1994, 1996) and *DEJAVU* (Kleywegt & Jones, 1997*c*)] that will compare a model to a database of known structures to try and find similarities. Sometimes unexpected similarities are found which in themselves may provide further insight into the function and/or evolution of the proteins involved.

Sometimes protein structures show similarities on a much smaller level than that of the fold or even domain structure, *e.g.* involving a limited number of side chains. This problem of three-dimensional pattern recognition in structures was discussed by Lesk as early as 1979 (Lesk, 1979). Pharmacophoric pattern matching is a well known technique in the context of chemical structure databases (Willett, 1987). We have developed a set of

programs that aid in identifying local similarities in macromolecular structures, inspired by the work of Artymiuk *et al.* (1994). *SPASM* (Kleywegt, 1998) is a program that can be used to find out if a user-defined motif also occurs in any previously solved structures. A motif is defined as a (usually small) set of residues for each of which the main chain and/or side chain must be matched in database proteins. A motif may be any constellation of residues that the user deems interesting, *e.g.* a hydrophobic cluster, a catalytic triad, a binding site for an inorganic ion, ligand, substrate or co-factor, or simply an unusual loop or interaction between two or



(*a*)



(*b*)

Fig. 8. Illustration of the use of *SPASM*. (*a*) A search for loops similar in conformation to residues 98–106 in cellular retinoic acid-binding protein type II (PDB code 1CBS). *SPASM* finds a number of hits, and has generated a macro file for *O* which has automatically read, superimposed and drawn these hits onto the target loop. (*b*) A search for histidine-triads similar to that observed in the structure of nitrite reductase, where it binds copper (PDB code 1AFN). Similar motifs are found in the structures of ascorbate oxidase, adenosine deaminase, carbonic anhydrase, haemocyanin and ecotin.

more residues. In the database, a residue's main chain is represented only by its Cα atom, and its side chain by the centre-of-gravity of all its side-chain atoms. This makes the database screening very fast, and enables the use of 'fuzzy patterns' (see below). The current *SPASM* database contains 2190 structures from the PDB (June 1998 release) whose sequences are mutually less than 95% identical (Hobohm & Sander, 1994). If a protein is encountered that contains a similar constellation of residues as that defined by the user, instructions are written to a macro file for *O*. When this macro is executed, all hits will be retrieved and superimposed onto the user's model. In order to allow 'fuzzy pattern matching', an option has been included to allow variations of the user-defined motif [namely, conservative substitutions as defined by the BLOSUM-45 substitution matrix (Henikoff & Henikoff, 1992)]. The program has been used in the analysis of an unusual Met–Trp interaction in the interface of the complex between acetylcholinesterase and the snake toxin fasciculin (Harel *et al.*, 1995), a set of five carboxylate residues important for the structure and function of inorganic pyrophosphatase (Heikinheimo *et al.*, 1996), and the P-loop phosphate-binding motif of phosphoenolpyruvate carboxykinase (Matte *et al.*, 1996). Two additional examples are shown in Fig. 8, and others are discussed by Kleywegt (1998).

RIGOR (Kleywegt, 1998) is a program that does essentially the opposite of *SPASM*. *RIGOR* uses a database of pre-defined motifs and scans the user's model to find out if any of these motifs occur in it. This program, in other words, is more or less a three-dimensional equivalent of PROSITE (Bairoch & Bucher, 1994). The generation and annotation of a high-quality database of pre-defined motifs is a major undertaking, that should preferably be coordinated by a database centre. For the time being, a motif database is used that is generated automatically by a program (called *AUTOMOTIF*) that looks for 'interesting' constellations of residues, such as hydrophobic clusters, charged clusters, mixed clusters, and sets of residues in the proximity of a hetero entity (ligand, ion, substrate, *etc.*). This motif database contains ~3400 entries at present.

Finally, structural biologists can take their models and attempt to do some 'database mining' in sequence, rather than structure, databases. If two or more models with structural similarities are available, their structure-based sequence alignment can be useful in efforts to identify other proteins (whose structure has not been determined yet) that may have a similar structure and/or function. *STRUPAT* (GJK, unpublished results) is a program that generates PROSITE-style sequence patterns [such as 'G-X-(WY)'] on the basis of a set of structurally aligned protein models. It considers only those parts of the models that are structurally similar in each of the aligned models, and identifies common residue types. These PROSITE-style pattern(s) can then be scanned against the SWISS-PROT and TrEMBL databases to identify other proteins that also contain the pattern.

An even more powerful means of identifying proteins with weak sequence similarities is based on the use of sequence profiles (Gribskov *et al.*, 1987, 1990; Gribskov & Veretnik, 1996). A profile is usually based upon a multiple sequence alignment. It attaches a score to each of the 20 amino-acid types (as well as one for gap opening and extension, respectively) for each of the residues in a sequence. Conserved residues will lead to very high scores for one or more residue types, and lower scores for all others, whereas variable positions will tolerate more diverse residues. *STRUPRO* (GJK, unpublished results) generates profiles based on aligned structures, again only considering the structurally conserved regions and ignoring the parts in between. In the profiles produced by this program, insertions or deletions inside the structurally conserved stretches are highly penalized, whereas insertions may be made between them with impunity. These profiles can subsequently be used to scan the SWISS-PROT database to identify other proteins that may have a similar (domain) fold (and, perhaps, a related function), even though the sequence similarities may be weak.

9. Outlook

We have attempted to illustrate that, during the past ten years, structural databases (both 'raw' and derived ones) have become indispensable tools at many stages of the process of protein structure determination, validation, and analysis. This trend is only likely to be amplified in the future.

(*a*) For model-building and validation purposes, the increasing number and structural variety of atomic resolution protein models will enable us to generate more comprehensive and reliable databases (*e.g.*, main-chain and rotamer databases), improved geometric and stereochemical restraints, and improved statistics concerning protein structure (*e.g.*, core Ramachandran areas).

(*b*) The structural database may also be of use in the future in the actual structure solution process. For example, as more and more (domain) folds are known, automated molecular replacement calculations may well become feasible. In cases where phase information is available, this can be used to improve the results of such computations (Kleywegt & Jones, 1997*a*). Finally, we have demonstrated previously (Kleywegt & Jones, 1994, 1997*c*) how fold-recognition techniques can be employed at the stage where (some) secondary-structure elements are visible in the skeleton, even though their directionality and connectivity may still be unknown.

(*c*) As the size of the structural database increases, it will become almost impossible for most structural biologists to memorize the details of each and every protein structure ever determined. Hence, derived databases for recognizing folds and smaller structural motifs will become indispensable tools. In favourable cases, such analyses may provide unexpected links between functionally diverse proteins, which in turn can be used to analyse the protein sequence databases.

(*d*) Genome sequencing efforts the world over are now producing a veritable deluge of sequence information. Functional genomics (*i.e.*, the study of the function of expressed genes whose sequences do not resemble those of other characterized proteins) is only just starting as a discipline, and 'high-throughput structural biology' is in its infancy. Structural databases are likely to play an important role in the process of predicting structure and function of the products of newly sequenced genes. In addition, a rapidly expanding structural database will also be beneficial in the area of homology modelling and that of fold recognition.

This work was supported by the Swedish Foundation for Strategic Research (GJK), its Structural Biology Network (GJK), the European Union (TAJ), and Uppsala University (TAJ). We thank Dr Christina Divne (Uppsala) for providing the coordinates of the first two models of cellobiohydrolase I, and Professor Axel T. Brünger (Yale University) for allowing us to use a pre-release version of the *CNS* software system.

References

Aalten, D. M. F. van, Bywater, R., Findlay, J. B. C., Hendlich, M., Hooft, R. W. W. & Vriend, G. (1996). *J. Comput. Aided Mol. Design*, **10**, 255–262.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans. II*, S1–S19.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. & Willett, P. (1994). *J. Mol. Biol.* **243**, 327–344.

Bairoch, A. & Apweiler, R. (1997). *Nucleic Acids Res.* **25**, 31–36.

Bairoch, A. & Bucher, P. (1994). *Nucleic Acids Res.* **22**, 3583–3589.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.

Brünger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L.M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Denson, D. A., Boguski, M. S., Lipman, D. J. & Ostell, J. (1997). *Nucleic Acids Res.* **25**, 1–6.

Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). *Science*, **265**, 524–528.

Editorial (1997). *Nature Struct. Biol.* **4**, 961–964.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Greer, J. (1981). *J. Mol. Biol.* **153**, 1027–1042.

Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.

Gribskov, M., Lüthy, R. & Eisenberg, D. (1990). *Methods Enzymol.* **183**, 146–159.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Gribskov, M. & Veretnik, S. (1996). *Methods Enzymol.* **266**, 198–212.

Harel, M., Kleywegt, G. J., Ravelli, R. B. G., Silman, I. & Sussman, J. L. (1995). *Structure*, **3**, 1355–1366.

Heikinheimo, P., Lehtonen, J., Baykov, A., Lahti, R., Cooperman, B. S. & Goldman, A. (1996). *Structure*, **4**, 1491–1508.

Hendrickson, W. A. & Konnert, J. H. (1980). In *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.25. Bangalore: Indian Academy of Science.

Henikoff, S. & Henikoff, J. G. (1992). *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Herzberg, O. & Moult, J. (1991). *Proteins Struct. Funct. Genet.* **11**, 223–229.

Hobohm, U. & Sander, C. (1994). *Protein Sci.* **3**, 522–524.

Holm, L. & Sander, C. (1994). *Proteins Struct. Funct. Genet.* **19**, 165–173.

Holm, L. & Sander, C. (1996). *Methods Enzymol.* **266**, 653–662.

Hooft, R. W., Sander, C. & Vriend, G. (1997). *Comput. Appl. Biosci.* **13**, 425–430.

James, M. N. G. & Sielecki, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). *J. Mol. Biol.* **125**, 357–386.

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

Jones, T. A. (1985). *Methods Enzymol.* **115**, 157–171.

Jones, T. A. & Kjeldgaard, M. (1994). In *From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 1–13. Warrington: Daresbury Laboratory.

Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–198.

Jones, T. A., Kleywegt, G. J. & Brünger, A. T. (1996). *Nature (London)*, **383**, 18–19.

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.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Nature (London)*, **185**, 422–427.

Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.

Kleywegt, G. J. (1997). *J. Mol. Biol.* **273**, 371–376.

Kleywegt, G. J. (1998). Submitted.

Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.
Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 858–863.
Kleywegt, G. J. & Jones, T. A. (1994). In *From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 59–66. Warrington: Daresbury Laboratory.
Kleywegt, G. J. & Jones, T. A. (1995*a*). *Structure*, **3**, 535–540.
Kleywegt, G. J. & Jones, T. A. (1995*b*). In *Making the Most of your Model*, edited by W. N. Hunter, J. M. Thornton & S. Bailey, pp. 11–24. Warrington: Daresbury Laboratory.
Kleywegt, G. J. & Jones, T. A. (1996*a*). *Acta Cryst.* D**52**, 829–832.
Kleywegt, G. J. & Jones, T. A. (1996*b*). *Structure*, **4**, 1395–1400.
Kleywegt, G. J. & Jones, T. A. (1997*a*). *Acta Cryst.* D**53**, 179–185.
Kleywegt, G. J. & Jones, T. A. (1997*b*). *Methods Enzymol.* **277**, 208–230.
Kleywegt, G. J. & Jones, T. A. (1997*c*). *Methods Enzymol.* **277**, 525–545.
Kleywegt, G. J., Zou, J.Y., Divne, C., Davies, G. J., Sinning, I., Ståhlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T. T. & Jones, T. A. (1997). *J. Mol. Biol.* **272**, 383–397.
Kraulis, P. J. & Jones, T. A. (1987). *Proteins Struct. Funct. Genet.* **2**, 188–201.
Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1996). *Protein Sci.* **5**, 1067–1080.
Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1997). *J. Magn. Reson.* **125**, 171–177.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *J. Mol. Biol.* **231**, 1049–1067.
Lesk, A. M. (1979). *Commun. ACM*, **22**, 219–224.
MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). *Curr. Opin. Struct. Biol.* **4**, 731–737.
Matte, A., Goldie, H., Sweet, R. M., & Delbaere, L. T. J. (1996). *J. Mol. Biol.* **256**, 126–143.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins Struct. Funct. Genet.* **12**, 345–364.
Noble, M. E. M., Zeelen, J. P., Wieringa, R. K., Mainfroid, V., Goraj, K., Gohimont, A. C. & Martial, J. A. (1993). *Acta Cryst.* D**49**, 403–417.
Oldfield, T. J. & Hubbard, R. E. (1994). *Proteins Struct. Funct. Genet.* **18**, 324–337.
Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B. W., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J. S., Verschueren, K. H. G. & Goldman, A. (1992). *Protein Eng.* **5**, 197–211.
Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.
Pearson, W. R. & Lipman, D. J. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Phillips, D. C. (1970). *British Biochemistry, Past and Present*, edited by T. W. Goodwin, pp. 11–28. London: Academic Press.
Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.
Priestle, J. P. (1994). *Structure*, **2**, 911–913.
Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.
Ramakrishnan, C. & Ramachandran, G. N. (1965). *Biophys. J.* **5**, 909–933.
Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 327–344.
Sonnhammer, E. L. L. & Kahn, D. (1994). *Protein Sci.* **3**, 482–492.
Vriend, G. & Sander, C. (1993). *J. Appl. Cryst.* **26**, 47–60.
Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Acta Cryst.* B**29**, 943–956.
Willett, P. (1987). *J. Chemometrics*, **1**, 139–155.
Zou, J. Y. & Mowbray, S. L. (1994). *Acta Cryst.* D**50**, 237–249.