Acta Cryst. (1998). D**54**, 1085–1094

Macromolecular Structure Databases: Past Progress and Future Challenges

Helge Weissig,[a,b]* Ilya N. Shindyalov[a] and Philip E. Bourne[a,b,c]

[a]*San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA,* [b]*Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, and* [c]*The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.*
E-mail: bourne@sdsc.edu

(*Received 17 April 1998; accepted 22 July 1998*)

Abstract

Databases containing macromolecular structure data provide a crystallographer with important tools for use in solving, refining and understanding the functional significance of their protein structures. Given this importance, this paper briefly summarizes past progress by outlining the features of the significant number of relevant databases developed to date. One recent database, PDB+, containing all current and obsolete structures deposited with the Protein Data Bank (PDB) is discussed in more detail. PDB+ has been used to analyze the self-consistency of the current (1 January 1998) corpus of over 7000 structures. A summary of those findings is presented (a full discussion will appear elsewhere) in the form of global and temporal trends within the data. These trends indicate that challenges exist if crystallographers are to provide the community with complete and consistent structural results in the future. It is argued that better information management practices are required to meet these challenges.

1. Introduction

Databases derived from the over 7000 files (1 January 1998) of macromolecular structure data comprising the Protein Data Bank (PDB; Bernstein *et al.*, 1977) are important to the macromolecular crystallographer who is both a depositor and user of the data. This importance increases proportionally to the contents of the PDB, which is growing at a near exponential rate, promising over 20 000 structures by the end of the year 2000. Databases containing data from the PDB are essential to efficiently explore this large corpus. From a crystallographer's perspective this exploration seeks to answer a number of questions. Has the structure in which I am interested been solved already? Is this protein fold I have determined unique? How should I classify this new protein structure? From a user's perspective a multitude of questions are being asked about structure–function relationships, both with respect to a single structure and multiple structures from the same family of proteins. Creating the infrastructure to answer these questions defines the relatively new field of structural bioinformatics. A major challenge for this new field is to effectively use data already generated and to insure that future data are collected in a way that facilitates direct loading into the necessary databases.

On one hand, the current situation is gratifying – there is so many valuable structure data available for exploring the function of biological macromolecules. On the other hand, the situation is worrying for it calls for an appraisal of the quality, accessibility and level of detail found in the corpus. This paper briefly addresses these contrasting situations. First, by giving an overview of past progress in developing macromolecular structure databases, and second, by using a specific database to examine the current contents of the corpus. Neither this review of past progress, nor a critical look at the current corpus, is new. The review of past progress should be considered an introduction and an update to a paper on structure databases that appeared two years ago (Gray *et al.*, 1996). Similarly, much as been written about the stereochemical quality of entries in the corpus (*e.g.*, Morris *et al.*, 1992; Laskowski *et al.*, 1993), and errors (*e.g.*, Kleywegt & Jones, 1995). This paper supplements those observations through a look at the current level of machine-readable annotation and inconsistencies in reporting structure information. We conclude that changes are needed in the current methods of data collection and curation if we are to make better use of macromolecular structure data *en masse*.

1.1. *Databases 101*

Why do we need structure databases? In simple terms, a database is software that permits you to organize the data to efficiently answer a specific set of questions of that data. For example, in finding all structures in the PDB that belong to the protein kinase family it does not make sense to search the more than 3 Gbytes of text information in the current PDB looking for the text string 'protein kinase'. It is more efficient to have previously partitioned the data, so for example, you only search the equivalent of all PDB COMPND records searching for references to the text string 'protein

kinase.' However, once you have found an instance of 'protein kinase', how do you reference other features of that structure, for example, the resolution? In any database associated with the instance of 'protein kinase' is a unique identifier for that protein. That same identifier is associated with other features of the same protein, including the resolution and is used to locate that feature. Similarly, by grouping the resolutions of all protein structures together makes returning all structures of better than, or less than, a certain resolution very efficient. There are three major types of database that implement this generic idea of partitioning data and all have been used for storing and querying macromolecular structure data. They are introduced in increasing order of complexity. The *indexed file* uses the basic file structure of the computer operating system along with an index as a unique identifier to partition the data. The *relational database* partitions data into tables where a tuple is a row of the table having a unique identifier (the primary key) for one or more attributes. So the PDB identifier could be a primary key and the compound name an attribute. The *object-oriented database* partitions data into classes where each tuple is an object instance from that class with an object identifier.

Each of these three database types has positive and negative features. Object-oriented databases can contain complex objects (*e.g.*, the surface of the active-site region of a protein) and have methods associated with each object instance. The latter implies that a query can also perform a calculation on data without being passed to an external program. Conversely, object-oriented databases are relatively inefficient and difficult to develop. Relational databases are more efficient but more limiting in the type of data they can contain. Indexed files are simple to implement, but have none of the characteristics associated with a formal relational database, such as concurrency control, enhanced security and formal query language. It is beyond the scope of this paper to discuss these issues in detail. Interested readers can refer to Bourne *et al.* (1998) for a discussion aimed at crystallographers wishing to access data from multiple databases, and Date (1994) for a detailed treatment of databases.

2. Past progress

Macromolecular structure databases date back to the work of Todd *et al.* (1984) who developed a relational database to be used with the graphical display of structure. Later Rawlings (1988) constructed a database to search for secondary-structure motifs. The first more generally applicable database was BIPED developed by Islam & Sternberg (1989) and this later, with the work of Thornton and Gardner, became a commercial product *Iditis* distributed by Oxford Molecular Inc. (Gardner & Thornton, 1998). BIPED was a relational database capable of asking a variety of questions relating primarily to geometry, for example, questions about hydrogen bonds, disulfide bridges, and torsional angles. The relational database SESAM (Huysmans *et al.*, 1991) followed, and included various energy parameters used in modeling and conformational energy calculations. The first object-oriented database was that of Gray *et al.* (1990) which was later used to search for hydrophobic subdomains (Kemp & Gray, 1990). Berman and colleagues developed the Nucleic Acid Database (NDB; Berman *et al.*, 1992) containing many unique derived data specific to nucleic acids. Our own work provided a general purpose C++ class library for representing protein structure for use in a variety of programs (Chang *et al.*, 1994). This was made persistent and accessible through the World Wide Web (WWW) in the MOOSE database (Shindyalov *et al.*, 1995) specializing in property pattern searching, where properties are represented linearly, that is, with respect to the primary sequence, and not spatially. A special purpose query language was also developed to ask complex questions of this and other databases (Shindyalov *et al.*, 1994).

Specialized centers began to support a variety of databases. The European Molecular Biology Laboratory (EMBL) provided such databases as HSSP [structure–sequence alignments (Schneider & Sander, 1996)] and FSSP [structure–structure alignments (Holm & Sander, 1997)], now located at the European Bioinformatics Institute. ExPASy provides SWISS-PROT, The National Center for Biotechnology Information (NCBI) provided the Macromolecular Modeling Database (Hogue *et al.*, 1996) based on the earlier work of Bryant using the statistical language S (Bryant, 1989). With the wide availability of the WWW, the intricacies of individual databases were hidden behind similar interfaces accessible through a WWW browser. Further, a small measure of database interoperability was achieved through hyperlinks between items of data within disparate data sources. As pointed out in Bourne *et al.* (1998), this form of interoperability is limited yet nevertheless useful. Higher levels of interoperability are achieved with systems such as SRS (Etzold *et al.*, 1996) and Entrez (Schuler *et al.*, 1996) which reformat the various forms of data, sequence, structure, bibliographic information, into a single database system. It is interesting that these systems use the index-file approach, which is the least sophisticated of the database types. We have recently generalized this approach in what is called the Property Object Model (POM; Shindyalov & Bourne, 1997). The POM approach is used for PDB+, the database used for the study described here.

With the rapid expansion of available structure data, a number of investigators have addressed the comparative analysis of proteins both from a structural and functional point of view. This has led to several protein-classification schemes, for example, the Structural Classification of Proteins (SCOP; Hubbard *et al.*, 1997)

Table 1. *A supplementary list of databases to that provided by Gray et al. (1996)*

Name (and function)	URL	Reference
The Protein Kinase Resource	http://www.sdsc.edu/kinases	Smith *et al.* (1997)
PDBObs (obsolete PDB entries)	http://pdbobs.sdsc.edu	Weissig & Bourne (1998)
EF-hand Calcium Binding Proteins	http://chazin.scripps.edu/cabp_database/	Nelson, unpublished work
G-coupled Protein Receptors	http://www.gcrdb.uthscsa.edu/	Kolakowski & Zhuang, unpublished work
HIV Proteases	http://www-fbsc.ncifcrf.gov/HIVdb/	Vondrasek & van Buskirk, unpublished work
Lipid Structures	http://www.lipidat.chemistry.ohio-state.edu/	Caffrey, unpublished work
Biological Relevant Structures	http://www2.ebi.ac.uk/msd/mm_search.shtml	Hendrick, unpublished work
Olderado (NMR ensembles)	http://neon.chem.le.ac.uk/olderado/	Sutcliffe, unpublished work
Peptidases	http://www.bi.bbsrc.ac.uk/Merops/MEROPS.HTM	Rawlings & Barrett, unpublished work
ReLiBase (receptor–ligand complexes)	http://www2.ebi.ac.uk:8081/home.html	M. Hendlich, unpublished work
Tops (Protein topologies)	http://www3.ebi.ac.uk/tops/	Westhead *et al.* (1998)
Esterases and Lipases (structure alignments)	http://cl.sdsc.edu/align_db.html	Bourne & Shindyalov, unpublished work

and protein class (C), architecture (A), topology (T) and homologous superfamily (H) [CATH (Orengo *et al.*, 1997].

We refer to the majority of macromolecular structure databases introduced thus far as broad but shallow. That is, they contain a limited amount of information for all known structures. A new type of database is emerging which is narrow but deep. These databases contain information on a subset of structural data in combination with other related information. The Protein Kinase Resource (PKR; Smith *et al.*, 1997) is one such resource which maintains sequence and structure alignments, disease-related information, lists of researchers, and so on, for a specific protein family. A number of similar resources can be reached from this site (Table 1).

In summary, there are a variety of databases providing information derived from macromolecular structure data that are accessible on the WWW. Table 1 provides a supplement to the list provided by the European Bridge Project (Gray *et al.*, 1996). In developing any of these databases, issues of data completeness, data consistency, data quality, and level of annotation arise. It is shown with some specific examples of queries from the PDB+ database that improvements are required in all of these areas.

3. Future challenges

The PDB+ database contains all macromolecular structures ever deposited with the PDB, that is, both current and obsolete structures. Thus, there are multiple versions of structures that have been deposited and subsequently replaced one or more times with newer versions by their authors. Contrast this to the PDB distribution which contains only the most recent version of a structure. There are currently 366 structures that are obsolete (1 April 1998). One obsolete entry may be replaced by multiple new entries and, conversely, several old entries can be replaced by a single new entry. The changes made when a structure is replaced are sometimes minimal and sometimes substantive. In most cases no annotation is provided within the replacing version to indicate why the previous version is being replaced. Simple typographic and other small syntactical corrections do not warrant replacing the entry. Access to obsolete PDB entries permits an analysis of how different versions of the same structure have changed over time. These changes reflect both advances in the field of structure determination and the correction of previous errors, and are thus valuable in improving our understanding of consistency within the corpus as it relates to time (temporal characteristics). The PDB+ query language also permits a series of questions to be asked of every structure ever submitted in order to determine overall (global) characteristics of this very valuable body of structure data. Answers to a few of these questions are given subsequently. A detailed discussion of the temporal and global characteristics of macromolecular data will be available elsewhere (Weissig & Bourne, 1998).

3.1. *Parsing PDB entries*

A major task in loading all PDB structures into any database is consistently parsing the complete PDB. Problems in parsing may not be apparent to a user of a single PDB file interested in only the atomic coordinates, but there are major problems when trying to consistently parse the complete corpus and extract information from the so-called ‘header records’, the information preceding the atomic coordinates. Morris *et al.* (1992) highlighted such problems and significant efforts have been made by the PDB and others to rectify these problems, but many problems remain. Problems such as undocumented inconsistencies between SEQRES records and the sequence found on ATOM records, the failure to use IUPAC notation, syntax errors, and inconsistent atom and group labeling. Note that these errors are distinct from the stereochemical quality of the structure but refer to the ancillary information so important for classifying structures according to name, taxonomy and experimental detail.

The problem with the current corpus begins with the format definition itself. The PDB format is defined in the *Contents Guide* provided by the PDB. However, the corpus contains entries in at least four distinct formats v1.0, v2.0, v2.1, and v2.2, with only those entries released from December 1996 indicating the format to which they comply through use of a REMARK 4 record. Early versions of the *Contents Guide* to which many structures conform are no longer available. Further, there is significant undocumented variation within a given format. For example, using PDB+ to analyze the change in the number of polypeptide chains from one version of a structure to another produces a number of surprises. While the application or removal of non-crystallographic restraints satisfactorily explained most of these surprises, several anomalies remained. Close inspection revealed that the definition of what constituted a polypeptide chain had changed from one version of a structure to another and single residues or non-polymer chemical components were assigned, or had removed, chain identifiers. This is obvious to an experienced human inspecting the entry, but not obvious to a computer program. While such a parser can be written to handle anomalies, it becomes a major software maintenance effort when the input format continues to change. Such anomalies are to be expected in an evolving discipline and are not unique to the representation of structure data, but are more frequent given the complexity of the information being represented. The use of a robust structure deposition tool (*e.g.*, *Autodep*) is a vital step in improving the consistency of entries. Further, current efforts by several groups to 'clean-up' existing PDB data by both removing anomalies and improving the level of annotation are most welcome, provided they lead to a single community-accessible version of each structure.

3.2. *Content of PDB+ and lack of experimental detail*

The analysis of PDB+ summarized subsequently raises a number of questions that remain unanswered because of missing or unparsable information relating to the experiment performed in determining the structure. Consider the case of resolution. Resolution is often used as the single yardstick for defining the accuracy of a structure and hence how that structure is used in a modeling experiment such as drug deign. However, there is variation in how the overall resolution is defined by different depositors. In principle 100% of the data in a given resolution range should be available for resolution to be reported at the limit of that range. This is often not the case. Version 2.0 and above of the PDB format provides REMARK 3 for including the completeness of the high-resolution shells, but it is not obligatory to provide this information. It should be possible to determine this information from the structure factors, but these are not available in many cases. It is particularly disappointing that experimental information as vital as that relating to resolution does not find its way into the final PDB deposition in a seamless and consistent manner, since it is available in electronic form at some point during the experiment. Closer interaction between software developers and the PDB could lead to better automated capture of experimental detail.

Consider another example of the lack of machine-readable experimental detail, which has important consequences when accessing structure information from a database, rather than simply reading it from PDB REMARK records. Interrogating the PDB+ database reveals that human deoxy hemoglobin (1HHB) was replaced by three entries (2HHB, 3HHB and 4HHB) in 1984, and these entries reside in the PDB today. A review of their fold deviation scores (FDS) color-coded and plotted with *Rasmol* (Sayle & Milner-White, 1995) on a per residue basis indicates three distinct situations (refer to http://pdbobs.sdsc.edu to review this case). The FDS is a useful first indicator when reviewing deviations from ideal geometry at each residue position. The FDS is the mean deviation from the ideal values of Engh & Huber (1991) summed over all bond lengths, bond angles, and dihedral angles in each amino-acid residue, including the side chain. The FDS plots reveal 2HHB to be a highly restrained model close to ideal geometry and 4HHB to be a loosely restrained model. 3HHB is presented as a highly restrained dimer to which non-crystallographic symmetry needs to be applied to generate the tetramer. Close inspection indicates that these three entries are derived from the same data set taken at 1.74 Å resolution and all three interpretations are correct within the accuracy of the experiment.

The following describes these observations in each of the PDB entries, 2HHB, 3HHB, and 4HHB

```
REMARK   6 THREE SETS OF COORDINATES FOR HUMAN HEMOGLOBIN WERE
REMARK   6 DEPOSITED SIMULTANEOUSLY.
REMARK   6  2HHB. REFINED BY THE METHOD OF JACK AND LEVITT.  THIS
REMARK   6      ENTRY PRESENTS THE BEST ESTIMATE OF THE
REMARK   6      COORDINATES.
REMARK   6  3HHB. SYMMETRY AVERAGED ABOUT THE (NON-CRYSTALLOGRAPHIC)
REMARK   6      MOLECULAR AXIS AND THEN RE-REGULARIZED BY THE
REMARK   6      ENERGY REFINEMENT METHOD OF LEVITT.  THIS ENTRY
REMARK   6      PRESENTS COORDINATES THAT ARE ADEQUATE FOR MOST
REMARK   6      PURPOSES, SUCH AS COMPARISON WITH OTHER STRUCTURES.
REMARK   6  4HHB. UNRESTRAINED REFINEMENT.  THIS ENTRY PRESENTS
REMARK   6      COORDINATES THAT ARE USEFUL FOR STATISTICAL STUDIES
REMARK   6      (E.G. PHI/PSI ANGLES) WHERE DATA UNBIASED BY
REMARK   6      RESTRAINTS IS REQUIRED.
```

While this is adequate for a human to decipher it is impossible to have a computer program, including a parser attempting to load these structures into a database, take this information into account when the guide in writing such a program states

REMARK 6 - 99, not assigned

Non-standard remark annotations or those with no clearly defined topic or assigned remark number appear with remark number 6 or greater, but less than remark number 100.

Non-standard remark annotations or those with no clearly defined topic or assigned remark number appear with remark number 6 or greater, but less than remark number 100.

Yet the distinction between these different atomic models can be important when using the data. For example, application of the Kabsch and Sander algorithm (Kabsch & Sander, 1983) for determining secondary structure provides slightly different answers using data from the three structures. While it can be argued that all determinations are correct within the accuracy of the original experiment and the user should be responsible for making themselves aware of such distinctions, this is just not practical when examining a large corpus. The level of machine-readable annotation that can be queried must be improved such that cases like that described above can be distinguished. Only then will future structure databases realise their full potential.

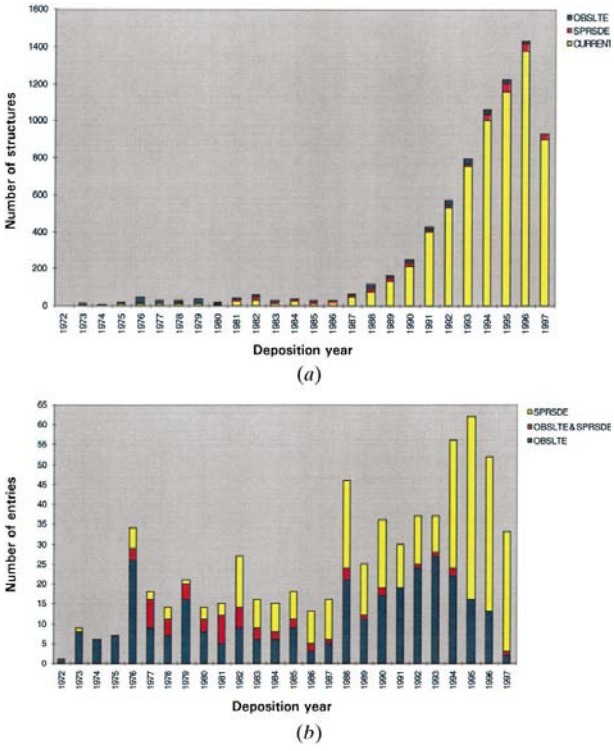

Fig. 1. Contents of PDB+ by year of deposition. OBSLTE indicates the entry is no longer in the current PDB distribution; SPRSDE indicates the entry has superceded another and is in the current distribution; CURRENT indicates the entry is part of the current distribution: (*a*) total number of deposited entries; (*b*) total number of obsolete and superceding entries.

3.3. *Analysis of the complete corpus*

Given the inconsistencies and lack of machine-readable annotation, what can be ascertained from the complete corpus? This question is addressed by queries of the PDB+ database, the results of which will be discussed. A more detailed set of queries and results is found in Weissig & Bourne (1998). It should be noted that we purposely examined the complete corpus, since this is what most users do, even though it is biased by a disproportionate numbers of certain types of structures. For example, there are currently over 600 lysozyme structures in the complete corpus, many determined at high resolution in the same laboratory.

4. Results

4.1. *The corpus*

The content of PDB+, given as the total number of depositions per year for the past 26 years since the PDB's inception, is shown in Fig. 1(*a*). Rapid growth began at the beginning of the decade, brought about with the insistence by many journals that structures be deposited with the PDB prior to publication and by the many advances in the field that have been documented previously (*e.g.*, Kleywegt & Jones, 1995). Throughout this 26-year history, the absolute number of replaced structures has remained relatively constant, but as a

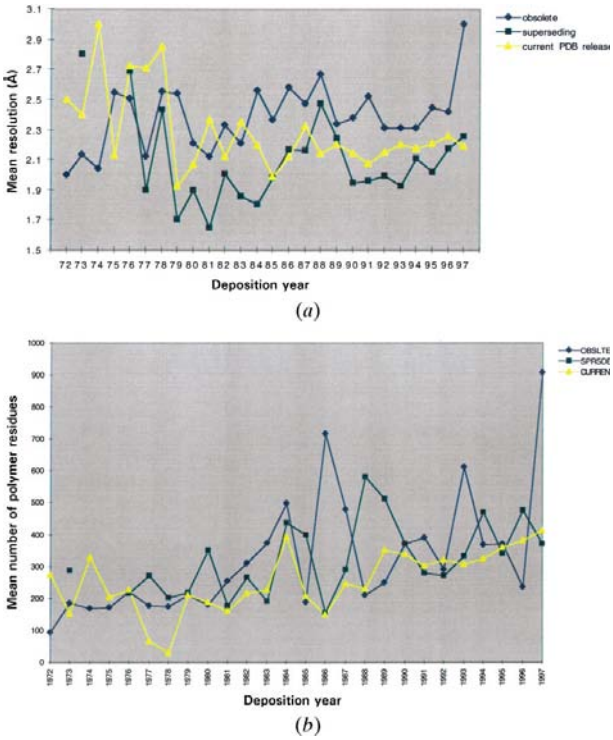

Fig. 2. Trends in PDB+ by year of deposition: (*a*) mean resolution (X-ray only); and (*b*) mean number of polymer atoms.

percentage of the total depositions in a given year it has decreased markedly over time. The majority of structures deposited before 1982 have been replaced at least once (Fig. 1*b*). All figures use data available as of 1 January 1998, however, it should be noted that data for 1997 are not complete since many depositions for that year were either still being processed or are on hold at the PDB. It is estimated that at least 600 structures

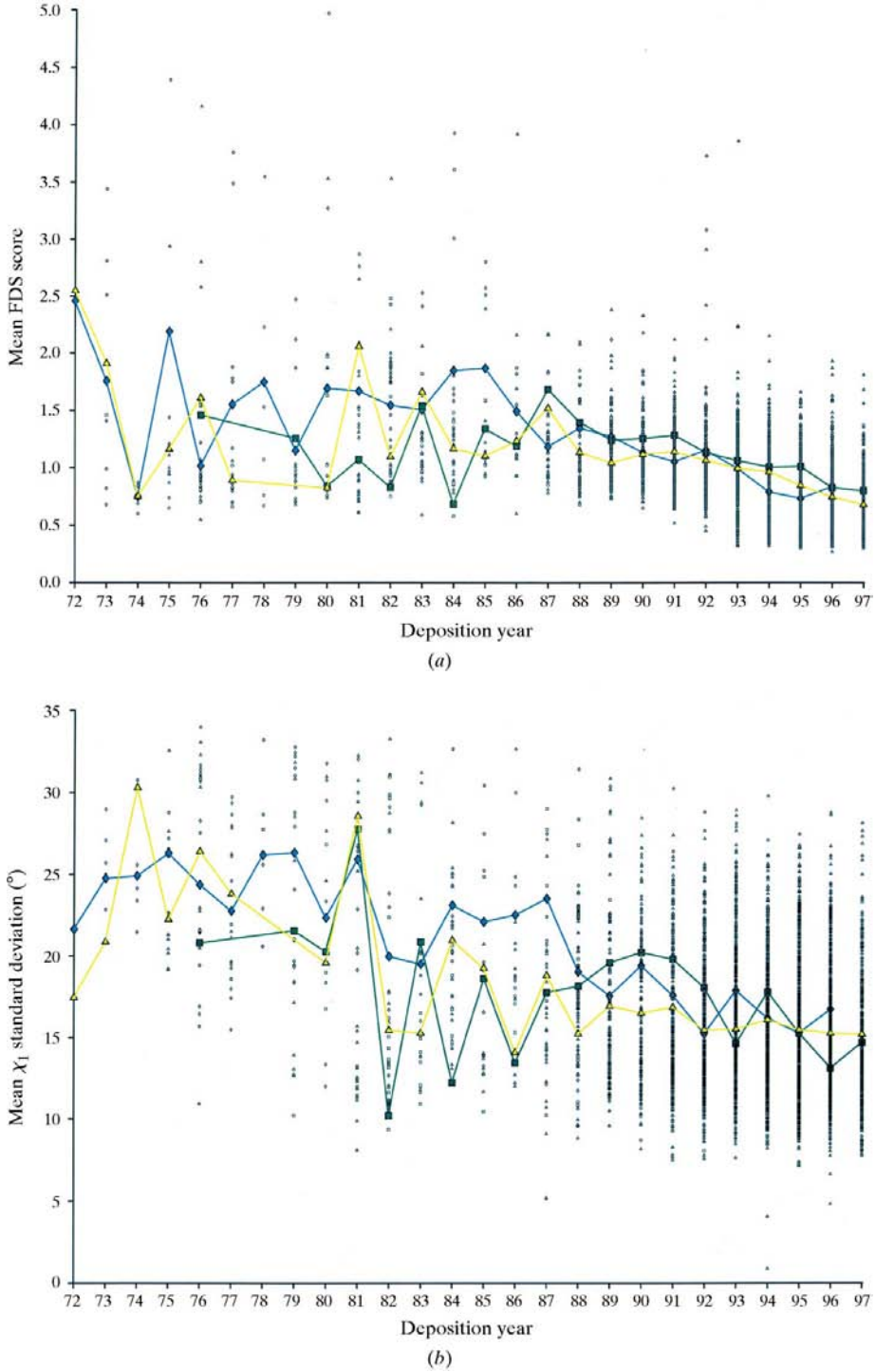

Fig. 3. Trends in PDB+ by year of deposition: (*a*) mean FDS score; and (*b*) χ for the current PDB (yellow), superceding entries (blue) and obsolete entries (green).

deposited in 1997 were not available as of 1 January 1998.

4.2. *Temporal trends*

Figs. 2(*a*) and 2(*b*) plot the mean resolution and mean size of structure, respectively, for all depositions made in a given year. This data varies widely in the early years when a small number of depositions were made per year, but since 1991, given a better statistical sample, both resolution and size show a slight upward trend. Since 1991 the mean resolution has increased from approximately 2.1 to 2.3 Å, while in the same time period the mean number of non-H atoms per structure has increased from approximately 2400 to 3000. This relationship is not surprising since complex structures with larger unit cells tend to diffract to a poorer resolution (see below). Thus, while the overall number of high-resolution structures deposited in a given year has increased over time (not shown), there is a predominance of large structures at slightly lower resolutions being deposited in successive years. This agrees with earlier findings using a set of 462 proteins (Morris *et al.*, 1992).

Fig. 3(*a*) plots the mean FDS for all structures deposited in a given year against the year of deposition. While the FDS is a crude measure, it is nonetheless clear that the overall stereochemical quality of structures deposited since 1990 has continued to converge towards ideal values. At the same time χ_1 values, which are frequently not restrained during refinement (Fig. 3*b*), have shown little change. One possible hypothesis that could be drawn from this analysis is that with the recent focus (beginning in 1990) on checking the stereochemical quality of models with programs like *PROCHECK* and *Whatif*, models are being over restrained during refinement to reach ideal values even when the data may dictate otherwise. Thus, rather than having a stereochemical model that matches the quality of the experimental data, the model matches an ideal representation of the structure. In principle this should lead to higher mean *R* values over time, but the reporting of *R* value is itself open to question (see below).

4.3. *Global trends*

The time of deposition is never a consideration when selecting a structure for further study. The most common yardsticks when seeking accurate data are the *R* value and resolution. Yet *R* value and resolution are poorly correlated (correlation coefficient 0.36) in the current

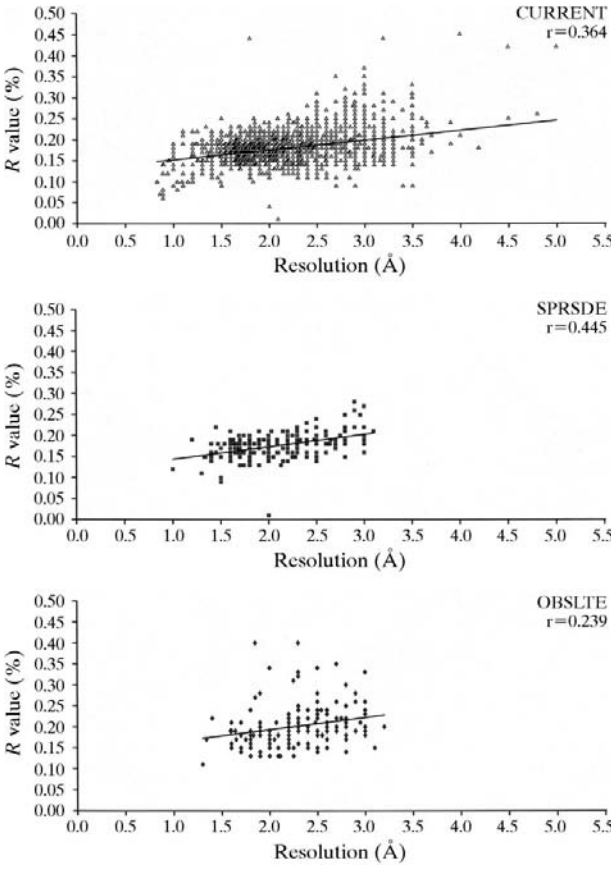

Fig. 4. *R* value (X-ray only) as a function of resolution for: (*a*) the current PDB; (*b*) superceding entries; and (*c*) obsolete entries.

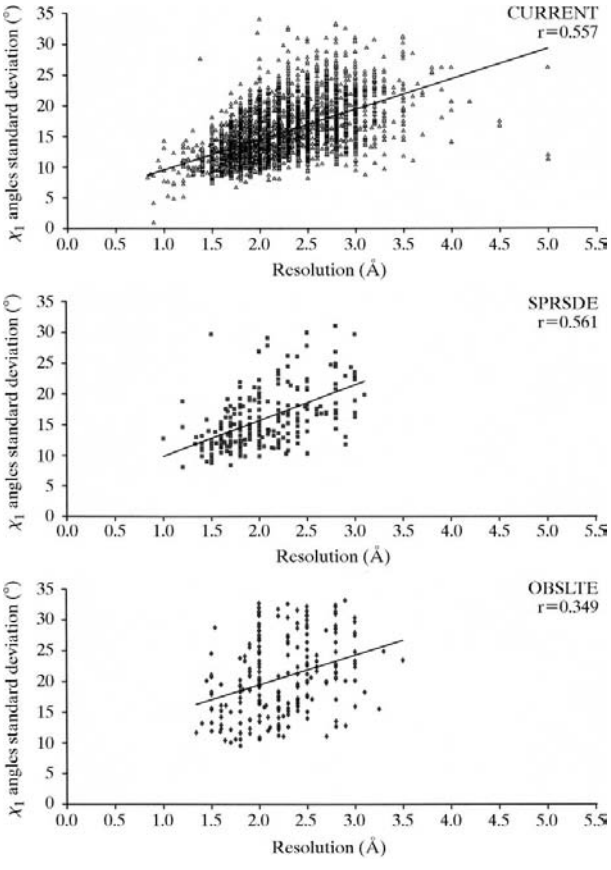

Fig. 5. χ_1 as a function resolution for: (*a*) the current PDB; (*b*) superceding entries; and (*c*) obsolete entries.

PDB as shown by the clustering in Fig. 4. That is, there are a significant numbers of high-resolution structures with poor R values and *vice versa*. As expected superceding structures show stronger clustering and correlation than the complete corpus and obsolete structures show less clustering. Outliers in these scatter plots can be explained in one of two ways. Either they are justifiable given the experimental conditions. For example, a high-resolution data set was collected but the quality of the diffraction spots was poor leading to a high R_{merge} and hence high R value. Or the outlier represents a potentially problematic structure. For example, an overrestrained model refined against a high-resolution data set. In either situation a user of that structure data should first be aware of this apparent anomaly and secondly should be able to probe the cause further by having experimental data available. Again χ_1 data are useful since they are less affected by the refinement. This is shown in Fig. 5 where there is a stronger correlation (correlation coefficient 0.58) between χ_1 and resolution. A more fundamental question to address is, what criteria are used in reporting data at a particular resolution? Early structures deposited with the PDB did not report the amount of data available in the high-resolution shells, or if they did it was buried in a non-parsable REMARK record. As of v2.0 of the PDB format this information is parsable, although the depositor has the option of whether to include it or not. For structures where the information is available, Fig. 6 plots the completeness of data in the high-resolution shell. There are a significant number of structures where less than 75% of the data have been collected at the reported resolution. A strong argument could be made that 100% of the data should be available at the reported resolution.

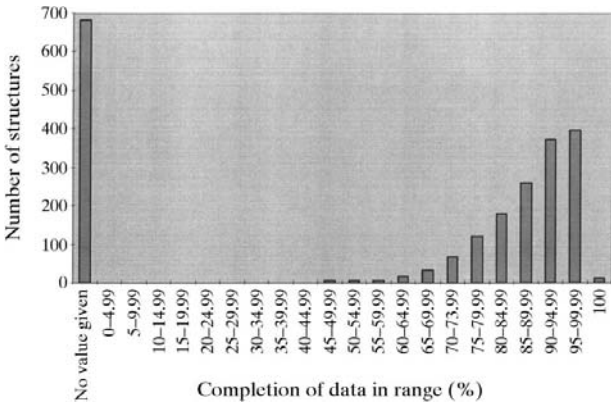

Fig. 6. Percent completeness of data in the high-resolution shell for all X-ray entries in PDB+.

There is evidence of inconsistent reporting of water content across the corpus, particularly in early versions of a structure. Looking at structures that have been replaced indicates that 73% reported increased water content, 12% no change, but 15% reported decreased water content, often from higher resolution data. This indicates over zealous reporting of water in the first structure determination. Brändén & Jones (1990) have concluded that for structures of 2 Å resolution or better one water molecule per residue is appropriate provided there is evidence of hydrogen bonding. Fig. 7 plots the average number of atoms and water molecules for all structures in PDB+ within 0.2 Å resolution increments. The correlation between number of atoms and resolution across the whole corpus is striking up to 2.6 Å. The overall trend in water content is as anticipated – less water per residue at low resolutions, more water per residue at higher resolutions. Specifically, between 3.0 and 3.2 Å there is approximately one water per 12 residues, between 2.2 and 2.4 Å one water per two residues, between 1.6 and 1.8 Å resolution, one water per residue, and between 1.0 and 1.2 Å resolution, three waters per residue. Thus, except at very high resolution, water is conservatively reported relative to the guidelines suggested by Brändén & Jones (1990).

The plateau in the mean number of waters per structure for structures solved between 1.6 and 2.4 Å resolution is noteworthy. Overall a constant number of

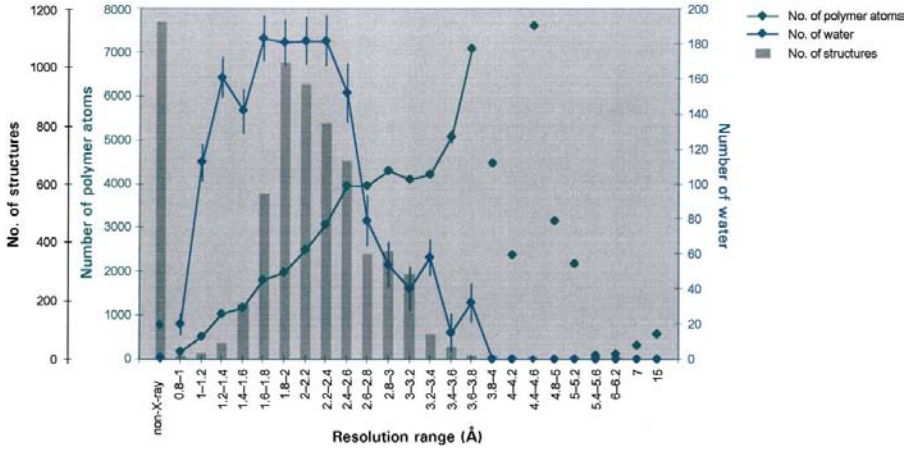

Fig. 7. Number of polymer atoms and water atoms as a function of resolution (X-ray only).

waters have been reported per structure regardless of resolution (and number of polymer atoms). While the number of polymer atoms ranged from 1700 to just over 3000, the mean amount of water remained constant at approximately 180 water molecules per structure. These findings need further study since they may be biased by a large number of very similar structures, for example the lysozymes, within this resolution range. Moreover, to be rigorous, water content should be plotted against the volume of exposed surface.

5. Discussion

PDB data available as either single structures, one per file, or databases either derived from all structure data, or subsets, have proved of immense value to the crystallographic and molecular biology communities. The value of these resources will increase further as their content grows. Nevertheless, several issues need to be addressed if these resources are to reach their full potential. The issues are data completeness (defined by what data can be read by computer), level of annotation beyond what is captured by the experiment, and consistency in reporting experimental details. The macromolecular crystallographic information file (mmCIF; Bourne *et al.*, 1997) offers one potential aid in solving at least some of these problems. Unlike the PDB format, mmCIF explicitly references every item of data by name and relationship to other items of data. That name is then defined in a formal dictionary (itself an mmCIF file) which can be used by the computer. In current parlance the mmCIF dictionary represents an ontology – a formal and extensive description of a particular field of study. Contrast this to the PDB *Contents Guide* that is less formal, less complete, and interpreted by humans and hence open to significantly different interpretations. Moreover, the dictionary provides the opportunity to validate each item of data item before it is accepted into the database since the dictionary includes such criteria as data type, allowable values (where applicable), and overall scope of data items to be included. In short, mmCIF or some other form of machine-readable data dictionary would lead to more complete and consistent reporting of structure details across the corpus.

Understandably crystallographers balk at the complexity of mmCIF, both in terms of the number of terms available in the dictionary and the complexity of the data representation. Much of this complexity can be hidden by the software used to collect user input in the data-deposition process, and by having more information collected by software in the structure-solution process automatically available as part of the deposition. The crystallographic community is currently addressing both of these solutions.

The first of the CCP4 study weekends was held in 1980, a time when protein crystallographers could easily recall all the major structures that had been solved. 18 years later that feat is no longer possible. We must now have a study weekend just to learn about the databases that are necessary for us to recall not just the structures that have been solved, but the characteristics of those structures and how they relate to each other. For those of us building these databases it is apparent that challenges exist if we are to capture consistent and complete data that make the databases as useful as possible. The time to meet those challenges is now, since the longer we delay amidst a large growth rate in the number of structures, the harder it will be to retroactively consistently annotate existing structures. The needed annotation is available either as part of the experiment or is present in the published paper. Modern information management practices make it possible to capture the necessary detail, and, as a service to the community, this should be performed.

References

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

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

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K., Westbrook, J. & Fitzgerald, P. M. D. (1997). *Methods Enzymol.* **277**, 571–590.

Bourne, P. E., Shindyalov, I. N., Smith, C. & Weissig, H. (1998). *Am. Trans.* In the press.

Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

Bryant, S. H. (1989). *Proteins*, **5**, 233–247.

Chang, W., Shindyalov, I. N., Pu, C. & Bourne, P. E. (1994). *Comput. Appl. Biosci.* **10**, 575–586.

Date, C. J. (1994). *An Introduction to Database Systems*. Reading, MA, USA: Addison-Wesley.

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

Etzold, T., Ulyanov, A. & Argos, P. (1996). *Methods Enzymol.* **266**, 114–128.

Gardner, S. & Thornton, J. (1998). *Acta Cryst.* D**54**, 1071–1077.

Gray, P. M., Kemp, G. J., Rawlings, C. J., Brown, N. P., Sander, C., Thornton, J. M., Orengo, C. M., Wodak, S. J. & Richelle, J. (1996). *Trends Biochem. Sci.* **21**, 251–256.

Gray, P. M., Paton, N. W., Kemp, G. J. & Fothergill, J. E. (1990). *Protein Eng.* **3**, 235–243.

Hogue, C. W., Ohkawa, H. & Bryant, S. H. (1996). *Trends Biochem. Sci.* **21**, 226–229.

Holm, L. & Sander C. (1997). *Nucleic Acids Res.* **25**, 231–234.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. & Choitha, C. (1997). *Nucleic Acids Res.* **25**, 236–239.

Huysmans, M., Richelle, J. & Wodak, S. J. (1991). *Proteins*, **11**, 59–76.

Islam, S. A. & Sternberg, M. J. (1989). *Protein Eng.* **2**, 431–442.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kemp, G. J. & Gray, P. M. (1990). *Comput. Appl. Biosci.* **6**, 357–363.

Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.

Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *J. Mol. Biol.* **231**, 1049–1067.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.

Orengo, C. A., Michie, A. D., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Rawlings, C. J. (1988). *Nature (London)*, **334**, 477.

Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.

Schneider, R. & Sander, C. (1996). *Nucleic Acids Res.* **24**, 201–205.

Schuler, G. D., Epstein, J. A., Ohkawa, H. & Kans, J. A. (1996). *Methods Enzymol.* **266**, 141–162.

Shindyalov, I. N. & Bourne, P. E. (1997). *Comput. Appl. Biosci.* **13**, 487–496.

Shindyalov, I. N., Chang, W., Pu, C. & Bourne, P. E. (1994). *Protein Eng.* **7**, 1311–1322.

Shindyalov, I. N., Cooper, J., Chang, W. & Bourne, P. E. (1995). *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, Maui. Los Alamitos, CA, USA: IEEE Computer Society Press.

Smith, C. M., Shindyalov, I. N., Veretnik, S., Gribskov, M., Taylor, S. S., Ten Eyck, L. F. & Bourne, P. E. (1997). *Trends Biochem. Sci.* **22**, 444–446.

Todd, S., Morffew, A. & Burridge, J. (1984). *Proceedings 3rd British National Conference on Databases*. Cambridge University Press.

Weissig, H. & Bourne, P. E. (1998). Submitted.

Westhead, D. R., Hatton, D. C. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 35–36.