Acta Cryst. (1998). D**54**, 1105–1108

Deposition of Macromolecular Structures

PETER A. KELLER,* KIM HENRICK, PHILIP MCNEIL, STUART MOODIE AND GEOFFREY J. BARTON

Macromolecular Structure Database Group, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England. E-mail: msd@ebi.ac.uk

(*Received 6 March 1998; accepted 18 June 1998*)

Abstract

Macromolecular structures are being determined at an increasing rate, and are of interest to a wide diversity of researchers. Depositing a macromolecular structure with the Protein Data Bank makes it readily available to the community. Accuracy, consistency and machine-readability of the data are essential, as are clear indications of quality, and sufficient information to allow non-experimentalists to interpret the data. Good-quality depositions are necessary to allow this to be achieved. The PDB's *AutoDep* system allows deposition and some preliminary automatic checking to take place at multiple sites, prior to full processing and release of the structure by the PDB. However, depositing a structure currently requires the manual entry of a large amount of information at the time of deposition. The data-harvesting approach will allow much more information to be deposited, without placing an additional burden on the depositor. Deposition-ready files will be generated automatically during the course of a structure-determination experiment. The additional information will allow improved validation procedures to be applied to the structures, and the data to be made more useful to the wider scientific community.

1. Introduction

When a macromolecular structure is released by the Protein Data Bank (Abola *et al.*, 1987; Bernstein *et al.*, 1977), it becomes immediately available to a wide community. A structure that is of particular scientific interest may be closely and critically examined by many people once it enters the public domain. A released structure has high visibility, since many of the structure-related resources that are available over the World Wide Web point to individual entries in the PDB. Deposition of a macromolecular structure can thus be considered to be a form of publication which compares in importance with publication in the traditional printed media. Even where a released structure has no conventional publication associated with it, the data are still readily accessible.

In recent years, there has been rapid growth in the numbers of biological macromolecular structures being determined and deposited. At the time of writing, approximately one quarter of the entries in the current PDB have been released during the previous 14 months. Structural information is also becoming more important to scientists who are not specialists in the experimental techniques of crystallography or nuclear magnetic resonance spectroscopy. These trends will require more, and higher quality, information on each structure to be made available. The ultimate source of most of this information is the data which is deposited by experimentalists. Accordingly, it is essential to develop new techniques to make deposition straightforward and less prone to error. Structural genomics research programmes that aim to determine large numbers of structures on a production-line basis, pose new challenges in terms of capturing, processing and releasing structural data.

2. Use of macromolecular structures

Before discussing the deposition of macromolecular structures, it is helpful to consider briefly how the information is used. This falls into two broad categories.

Firstly, the use of small numbers of structures by scientists who are interested in particular features of those structures. For this kind of study, accurate and detailed information about the structures studied is important. For such studies it is reasonable to examine each entry in detail and to consult the associated literature as appropriate.

Secondly, the use of the whole collection of structures, or large subsets of it. In these studies, it is rarely practical to examine individual entries in great detail. This kind of study often involves large numbers of comparisons to be made between structures. Normally, it is desirable for such comparisons to be automatic, in order to maintain consistency and minimize subjective judgements. This places two requirements on the information contained in the collection of structures: (i) it should be possible to interpret the information consistently across the entire collection, and (ii) the information should be machine readable, so that it can be reliably extracted by software.

For any kind of use it is desirable to have meaningful indications of the quality of the data. It is also important that the user is able to have an understanding of what

ISSN 0907-4449

the data represents, without needing detailed knowledge about the experimental techniques used for structural determination. Simple tasks such as the generation of a functional molecule by the application of crystal symmetry to the contents of the asymmetric unit, can be daunting for the non-specialist.

There are many resources available to the scientific community which provide access to the results of applying some form of analysis to the information contained in the Protein Data Bank. [Examples include CATH (Orengo *et al.*, 1997; see also http://www.Biochem.ucl.ac.uk/bsm/cath) and *3Dee* (Siddiqui & Barton, 1998; access to *3Dee* is *via* http://barton.ebi.ac.uk), which are discussed elsewhere in these proceedings.]† The ultimate source of these 'added-value' databases is information supplied by depositors of macromolecular structures, and to some extent, problems with the derived results reflect problems with the deposited information. For this reason (among others), it is important that depositions to the PDB are complete and accurate. In this context, it may be noted that if further information about a structure determination is required some time after deposition, obtaining this information can be difficult, if not impossible.

3. Current deposition procedure – AutoDep

Various methods of deposition to the PDB have been available, such as electronic mail or file transfer protocol (ftp). The principle underlying these methods is that at deposition time, all the relevant information is gathered together, and collated and entered by the depositor. This can be a time-consuming and error-prone process, particularly where the structure determination has taken several years and involved a large number of people.

Currently, the recommended method for deposition of macromolecular structures in the PDB, is *via* the World Wide Web *AutoDep* system.‡ This software, which was originally developed at the PDB, has been modified at the EBI to allow deposition to the PDB *via* other sites. (At the time of writing, the only site other than the PDB at BNL which is running *AutoDep* is the EBI.) *AutoDep* is a system that gathers the required information from the depositor. It is able to carry out basic checks on the items which are entered, such as ensuring that a date or numeric value conforms to syntactical rules. It also checks that the information supplied is complete, thus ensuring that it will be possible to prepare an entry from the original deposition. When all the necessary information has been supplied, the structure is validated with the program *WhatIf* (Vriend, 1990; http://swift.embl-heidelberg.de/whatif). The results of this validation are presented to the depositor, who can then decide to proceed with the submission, or alternatively to carry out more work on the structure and deposit a revised set of coordinates.

Up to this point, all processing is performed automatically at the site which is running *AutoDep*. When the depositor finally submits the structure, the deposited information is sent to the PDB who issue a PDB ID code. Further work on the entry beyond this point is carried out by PDB staff, who then release the fully annotated entry when processing is complete.

If a number of similar structures are to be deposited, then it is possible to base a new submission on a previous one, or on a released entry. While this is helpful for depositing a series of related structures, it is nevertheless important to make sure that inappropriate information is not carried over from one entry to another by mistake.

4. Data harvesting: deposition in the future

Much of the information that is required at deposition is calculated by the software used in the normal course of structure determination. The data-harvesting approach requires modification to key programs so that they produce 'deposition files'. These files will contain the information that is required for the submission of the structure, and will be retained by the experimentalist until the structure is ready to be submitted. They will then be uploaded along with coordinates and experimental data. The depositor will only be asked for information which is not contained in the files.

The deposition files are distinct from the normal output files (*i.e.* log and data files) that are produced by the software. Output file formats vary enormously from program to program, including programs that have the same general role in structure determination. Output formats often change from version to version of the software, and many experimentalists make modifications to suit their own purposes. The deposition files, on the other hand, will follow slowly changing specifications which are largely independent of the software, except that the information that they contain will reflect the function of the programs which generate them. Their contents will conform to the requirements of the database to which the information is to be deposited. It will not be necessary for depositors to examine the files, although they will of course be free to do so.§

The fundamental change of emphasis from current procedures, is that the required information is produced in a form which is ready to be deposited, *during* the

† For references to others, see for example, http://www.pdb.bnl.gov/pdb-docs/mole.html (or the corresponding link at any PDB mirror), http://www2.ebi.ac.uk/msd/msdlinks.shtml, or http://www.lmcp.jussieu.fr/sincris-top/themes/biologie/

‡ *AutoDep* is available from the following two URLs: http://www.pdb.bnl.gov:8080 and http://autodep.ebi.ac.uk.

§ Extensions to the mmCIF dictionary to define the contents of the deposition files are currently under review. For information on the status of these extensions, see the mmCIF section of the NDB Biological Structure Resource at http://ndbserver.rutgers.edu, or any of its mirror sites.

course of the experiment. Another way of looking at this, is to consider the deposition files to be snapshots of the current status at the end of each of the major stages of a structure-determination experiment.

AutoDep supports uploading of previously prepared files that follow the format of the header of a PDB entry, but this is somewhat limited. *X-PLOR* (Brünger, 1992) can generate such a file that contains some information. However, information relating to the processing of diffraction data is not available to *X-PLOR*, and so null values are written. Currently, no other software that is in widespread use in macromolecular crystallography or NMR writes files in this format. While this facility is undoubtedly in the spirit of the data-harvesting concept, it needs to be developed in two ways. Firstly, the procedure should be supported by a much wider range of software. Secondly, the format and contents of the deposition files will have a specification that is distinct from the PDB format. Breaking this link will allow greater flexibility in the range of information to be harvested, and the manner in which it is eventually presented to users of the information.

The outline of the projected data-harvesting process is shown in Fig. 1. At the lower left of the diagram, there is a category for information which must be entered manually. This category of information includes that which is not known to any software package, such as the source of a protein and the method of crystallization.

There is general agreement in principle between the MSD group at the EBI, the PDB project, and many of the developers of the relevant software, that the implementation of the data-harvesting approach should proceed.

5. Future directions for data on macromolecular structures

The impetus for developing an improved deposition process comes in part from the increasing rate of growth of the PDB, and the requirement to make its contents available in a meaningful way to scientists who are not members of the X-ray and NMR communities. The growth rate is of greatest concern to those involved in the curation, maintenance and delivery of the information. However, it is unlikely that the resources available to maintain the databases will increase in direct proportion to the quantity of data, so it is important that new methods are developed before the growth presents a serious problem.

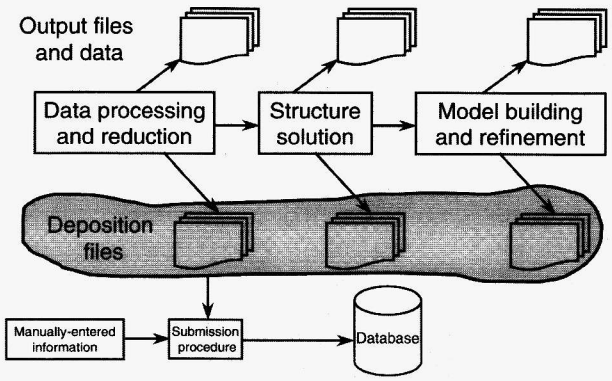

Fig. 1. Schematic outline of the data-harvesting process.

The detailed consequences of the broader scientific interest in structure are more difficult to anticipate. The data are likely to be used more routinely in the design of laboratory experiments. These uses are distinct from deriving information by analysis of structural data: making a decision on how to spend time and resources on the basis of features of one or more macromolecular structures, requires confidence that those features are not errors or artefacts of the structure-determination experiment. This is especially true where such features deviate significantly from what is normally observed, such as unusual peptide geometries. Where the structure has been retrieved from the PDB, the user of the information will not necessarily be in a position to communicate directly with anyone who was involved in the structure determination. Also, the user may be interested in aspects of the structure which were not of direct interest to the depositor. The salient question in such a case, is whether the experimental data confirm the presence of features in the structure which are of interest. Providing the answer is, in essence, validation, and implementing sophisticated validation protocols is a longer term aim in the provision of macromolecular structures to the scientific community. A discussion of validation is outside the scope of this article, but it is worth noting that the raw material for validation should be provided by the deposition process. This emphasizes once again the importance of ensuring that deposition provides complete and accurate information to the collection of macromolecular structures.

We wish to acknowledge the Wellcome Trust, the European Union and EMBL for their support. We are grateful to Professor Janet Thornton for invaluable help and encouragement in her rôle as scientific advisor to the MSD group. We also acknowledge valuable collaboration with the Protein Data Bank, the Nucleic Acid Database, and the National Center for Biotechnology Information.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank. Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography.

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

Brünger, A. T. (1992). *X-PLOR: A system for X-ray crystallography & NMR.* Yale University Press, New Haven, USA.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**(8), 1093–1108.

Siddiqui, S. A. & Barton, G. J. (1998). In preparation.

Vriend, G. (1990). *J. Mol. Graphics*, **8**, 52–56.