The Protein Data Bank and lessons in data management

Philip E. Bourne, John Westbrook and Helen M. Berman

Date received (in revised form): 17th December 2003

Philip E. Bourne
is a Professor of Pharmacology at UCSD, Director of Integrative Biosciences at SDSC, and a Co-Director of the PDB. His research interests focus on structural bioinformatics and high performance computing as applied to problems in genomics and proteomics.

John Westbrook
is a Research Associate Professor of Chemistry and Chemical Biology at Rutgers and a Co-Director of the PDB. His research interests include bioinformatics, computational biology and ontology management.

Helen M. Berman
is a Board of Governors Professor of Chemistry and Chemical Biology at Rutgers, and Director of the PDB. Her research interests include structural biology and bioinformatics, with a special focus on protein–nucleic acid interactions.

Keywords: *PDB, mmCIF, macromolecular structure data, the human factor, data uniformity*

Philip E. Bourne,
Research Collaboratory for Structural Bioinformatics,
Department of Pharmacology and San Diego Supercomputer Center,
University of California San Diego,
9500 Gilman Drive,
La Jolla,
CA 92093-0537, USA

Tel: +1 858 534 8301
Fax: +1 858 822 0873
E-mail: bourne@sdsc.edu

Abstract

The Protein Data Bank (PDB) is a widely used biological database of macromolecular structures with a long history. This history is treated as lessons learned and is used to highlight what are believed to be the best practices important to developers of biological databases today. While the focus is on data quality, data representation and the information technology to support these data, the non-data and technology issues cannot be ignored. The role of the human factor in the form of users, collaborators, scientific society and *ad hoc* committees is also included.

INTRODUCTION

The Protein Data Bank (PDB)[1,2] is the single worldwide repository for the structures of biological macromolecules and currently contains over 22,000 individual structure determinations derived experimentally by X-ray crystallography, nuclear magnetic resonance (NMR), neutron diffraction and, most recently, cryo-electron microscopy. These data have been accumulated over a 32 year period. The PDB provides a rich history from which to explore the practices of biological data management, as it contains a data set that has many characteristics found in other biological data – diversity, complexity, variable quantity and variable quality of annotation.

Biological data management concerns more than just the technical aspects; there are sociological and political issues as well. While our focus will be on the technical aspects of information management, the PDB has a human interface that has had an important impact on data availability and usage and is discussed. We begin with the scientific motivation for establishing the PDB.

THE SCIENCE PERSPECTIVE

In 1971, a small group of scientists interested in protein structure attending the Symposium on Quantitative Biology on Protein Crystallography[3] discussed the need to create an international and freely accessible repository for protein structure data. There were an estimated 15 protein structures at the time with the data stored in laboratory notebooks and on punch cards. It was decided that this repository be located at Brookhaven National Laboratory. By the time the standard reference for the PDB appeared in 1977[4] there were 77 structures in the archive. That reference has been cited over 6,500 times as of October 2003, indicating the wide usage that these data enjoy. A number of entries can be linked to Nobel Prizes, for example, myoglobin,[5,6] haemoglobin,[7] insulin[8] and the photoreaction centre.[9] The awards recognised the profound impact that those scientists have had in advancing our knowledge of living systems through the study of structure–function relationships. As the body of structural information grew, the notion of comparative structural analysis was born. This is exemplified by the early work on multiple haemoglobin structures which showed that fold was far more conserved than sequence,[10] spawning a whole field of comparative modelling. Today we have the emergent field of structural bioinformatics,[11] which encompasses various fields of study using all or a part of the structural corpus.

Examples are protein structure classification, protein structure prediction, the study of protein motion and the study of protein–protein interactions.

THE USER PERSPECTIVE

How and what specific PDB data are used reflects an increasingly diverse user base. What started as a small group of structural biologists and chemists has expanded to include computational chemists and bioinformaticians, biologists working at many different biological scales (genotype to phenotype), educators, students, artists and the public at large. If measured by access to the PDB primary web and ftp sites, this is an active community, with 10,000 people visiting the website each 24 hours and downloading, on average, one structure every second. Technically, the PDB is now coming to grips with the needs of this diverse user community through technology that permits us to present views of the data for specific user groups. This recognition is part of our future development effort and described subsequently.

10,000 people visiting the website every 24 hours

Another important component of the user community is that of researchers who deposit the data, in contrast to those researchers who only use the data. This distinction has diminished over time since more depositors have become users of other people's data, and the time required for structure determination has lessened. Nevertheless, depositors and users have specific needs and viewpoints that must be balanced. Structural biologists who deposit data maintain a sense of ownership and pride in these data, while users often view a given structure as a single data point. We believe that part of the success in maintaining an international data resource such as the PDB has been a management team that includes multiple perspectives, and who can work together to develop solutions that represent the best compromise to the needs of diverse communities.

Depositors and users have specific needs and viewpoints that must be balanced

Consider an example of this dynamic. A particular X-ray crystal structure might be difficult to crystallise. When small crystals are obtained after significant effort, they do not diffract well and lead to a poorly defined structure at low resolution. Nevertheless, the structure represents the best that could be obtained and leads to some new discoveries about the function of that protein. To a user interested in classifying protein folds this appears only as a low-quality data set. That is, the estimated net worth of a structure may be different in different user communities. The PDB's long-term approach to this dilemma is to provide each user community with the tools they need to make the best use of the data at hand. The PDB's role is not to 'police' the data. In this case, the PDB's role would be to highlight to biologists through links to functional annotation the role this structure plays in that understanding. To a bioinformaticist, the PDB's role would be to provide the tools to explore the details of that fold and tools that permit a determination of the stereochemical quality.

THE TECHNOLOGY PERSPECTIVE

The determination of macromolecular structures has been profoundly affected by technology. The advent of gene expression technologies, crystallisation robots, synchrotron radiation sources and high-frequency NMR are a few examples (see Markley *et al.*[12] for a review). We can expect more such developments in the current era of structural genomics[13] where technology development and improved engineering practices are an integral part of the initiative. Here we focus on the technology of data management and data distribution.

Data distribution has gone from taking days and months, depending on geographical location for postal delivery, to fractions of a second for web delivery for most users. From its beginnings of wide distribution on nine-track magnetic tape, the PDB has evolved to Internet services supported by the http and ftp protocols and most recently web services and CORBA for distributed access. Even

The PDB has evolved to Internet services supported by the http and ftp protocols and most recently web sevices and CORBA

with wide adoption of the internet, hard media are still needed for some users, and the media to support these distributions has changed several times and will continue to evolve. Associated with these developments are challenges in developing user interfaces for fast and efficient access by a variety of user types as outlined above. Today we face issues of usability engineering and the need to conduct focus groups to satisfy the customer, when in the early days a magnetic tape containing PDB and a printed readme file was enough.

Today we face issues of usability engineering

Hierarchical-, network-, relational- and object-oriented-based data models have all emerged since the PDB was founded, each with accompanying database management systems. These models have all been used in some form to manage PDB data with varying degrees of success.[11] In one sense, the data model is of secondary importance. Any data model is only as good as the data it represents. Good data need to be able to be parsed, loaded and represented with some measure of integrity. The need for good data and a good representation of that data so that a variety of data modelling techniques can be applied is perhaps the biggest lesson that can be learned in reading this paper. Data representation is arrived at by a process and is rarely the work of a single individual – which takes us to the human factor.

Any data model is only as good as the data it represents

THE HUMAN FACTOR

Several key events beyond the founding of the PDB have led to the PDB as it exists today. First and foremost is the community's decision to create a single repository for macromolecular structure data. This is rare in the world of biological data and has made it much simpler for scientists to work with these data than if they were distributed across multiple resources, as was true of the sequence databases at one time. This recognition of the importance of a single archive recently cumulated in the formation of wwPDB[14] – a consortium consisting of the Research Collaboratory for Structural Bioinformatics (RCSB), the Macromolecular Structure Database (MSD) group at the European Biotechnology Institute (EBI)[15] and PDBj at Osaka University in Japan,[16] which is dedicated to the maintenance of a single archive with the same definition even though the data are collected and processed by multiple sites.[17] Such agreements are not easy to achieve, yet very important to the users of the data.

The second human factor was the work of an *ad hoc* committee headed by Frederic Richards at Yale University and a committee set up by the International Union of Crystallography (IUCr), whose efforts led to a set of guidelines for the deposition and release of structures.[18] Both journals and government funding agencies adopted these guidelines, which required deposition prior to publication. This development was followed an immediate increase in the number of depositions to the PDB and created some challenges in data management which will be explored subsequently. This trend in data deposition has extended to the primary experimental data (X-ray diffraction intensities or NMR constraints) without a mandate from the journals or funding agencies. In 2003, approximately 71 per cent of structures deposited included primary experimental data.

In 2003, approximately 71 per cent of structures deposited included primary experimental data

A third human factor was the work of the mmCIF (macromolecular Crystallographic Information File) Working Group and the scientific society that encouraged that work. In 1990, the IUCr established a standard data representation referred to as the Crystallographic Information File (CIF).[19] The primary purpose of this data effort was to provide a common archive format for data deposited from a small molecule crystallography experiment. Deposition of these data were required as part of publication in IUCr journals. While initially it was not required that that data be deposited using the CIF format, the carrot offered was faster publication if that

format was used. In two years, over 90 per cent of small molecule data depositions were in CIF format, leading to a fast and wide adoption of this standard form of data representation. Hoping to build on the coat-tails of this success, the IUCr appointed the mmCIF working group to define a macromolecular structure version of this data representation by means of a few simple extensions. It was this step that made the PDB what it is today.

The PDB format has serious flaws as a rigorous data representation

DATA REPRESENTATION

The authors of the mmCIF dictionary were part of a working group chaired by Paula M. Fitzgerald of Merck Research Laboratories. What we anticipated would take a few weekends turned into several years before v1.0 of the mmCIF dictionary was officially released in 1997.[20] The gross underestimation in what a large task this would be resulted in part from content and part from context. Content addresses the question:

> How do I fully represent the complexity of the structure of a biological macromolecule and the experiment that determined its structure and do so in a way that is extensible so that anything unexpected that comes along in the future can also be fully represented?

The context issue addresses the question:

> How do I describe this complexity in a way that it can be fully utilised by computers with some concession to human readability?

(In retrospect, human readability can easily be achieved by a suitable tool, but we remind you not to underestimate the human factor at the time of this process – tools were not trusted to faithfully return the correct results.) The end result of this process is what is known today as the PDB exchange dictionary, which contains definitions for X-ray, NMR and cryo-electron microscopy experiments.[21] This dictionary permits data to be exchanged between sites processing PDB depositions and forms the basis of a rigorous data representation not possible from the original PDB format.[22]

The PDB exchange dictionary contains definitions for x-ray, NMR and cryoelectron microscopy experiments

The original PDB format was well suited to the early needs of data archiving, and, as a result, this format is recognised by all software that accesses macromolecular structure data today. However, it has serious flaws as a rigorous data representation from which complex databases with a fine level of detail can be built and queried. First, complex structures cannot be accommodated by the fixed field format. For example, the designation of a polypeptide chain and DNA strand is accommodated by a fixed length single character field. Some structures today, notably ribosomes, contain more such components than can be accommodated by using all available alphanumeric characters. Secondly, the limitations of a fixed length record based on a FORTRAN punched card and with poor atomicity in the data, that is, the lack of important items of data which can be discretely and consistently referenced, present problems in data parsing and loading of data into any type of database. mmCIF solves these problems and is the backbone of the PDB today.

mmCIF has been described fully elsewhere[23] and only a synopsis is given here. mmCIF conforms to a subset of encoding rules embodies in a Self-defining Text Archival and Retrieval (STAR) syntax.[24] STAR has provisions for defining scope, nesting, looping and so forth. Conforming to STAR is a Dictionary Definition Language (DDL) which defines how dictionaries are described. DDL has provisions for fully characterising the terms in the domain and is relational in nature. That is, there is the notion of relations (categories), attributes (specific data names), primary and secondary keys (mandatory data items) and so on. The data defined by mmCIF consist of name–value pairs where each name must be defined in the mmCIF dictionary. The mmCIF dictionary can be characterised as having the features of an extensible markup

The mmCIF dictionary can be considered an ontology for macromolecular structure

language (XML) document type definition or schema. It provides a specification for the content of a data files and information necessary to validate the data. The mmCIF dictionary can be considered an ontology for macromolecular structure and the experiment that derived that structure.[25]

With the mmCIF dictionary as the backbone of the PDB, the rest of the skeleton could be assembled. Each item of data described by the dictionary can have precise definitions and examples, controlled vocabulary, ranges of values, units of measure and so on associated with it. Thus, the dictionary can be used as a forms generator from which data can be entered. As the field of structure biology changes, new data items can be added to the dictionary and the forms interface regenerated automatically. A process overseen by the IUCr exists for vetting new data items to be included in this comprehensive and machine-readable view of a complex science.

The development of an ontology for the field of study is prerequisite to orderly and interoperable data resources

In retrospect, and in what we observe today in other work, the development of an ontology for the field of study is a prerequisite to orderly and interoperable data resources. The growing popularity of the Gene Ontology[26] is sample evidence for how this need is being met. Notwithstanding, mmCIF has not enjoyed wide popularity outside the PDB; users prefer to work with simpler PDB files. We believe this reluctance to be in part the complexity of mmCIF and the data being represented. In addition, it was not considered mainstream by computer scientists and software developers. Hence these groups were slow to develop software around mmCIF. Much of this perception has changed as the content of the mmCIF dictionary has been translated into XML schema and mmCIF data files into XML.[27]

A completely re-engineered system is in the alpha stage of development

THE PDB DATABASES

The exchange dictionary (described above as an extension to the mmCIF dictionary) is the conceptual schema from which physical databases can be instantiated. This is a key part of the PDB strategy now and going forward – the ability for users to establish their own databases of all or a subset of PDB data. Not only does it satisfy user needs, it also satisfies the partners that make up the wwPDB and who act as data deposition sites. While Osaka University and the Research Collaboratory for Structural Bioinformatics (RCSB) use the ADIT data deposition software, EBI uses Autodep software. Since data are collected at each site, the exchange dictionary provides consistency of data deposition. Having data collected independently helps with the unpredictable influx of data, provides different perspectives on the process, allows for periodic cross-checking of data consistency, and finally provides a global perspective to the resource. At the same time each site provides its own unique access to these data – a different physical instantiation in terms of database and views – but based on the same underlying data definition.

The current PDB production website available from the RCSB consists of a hybrid of databases – Sybase (relational database based on the mmCIF schema), POM (indexed records[28]), Lucene (indexed words and phrases[29]) and ASCII files. All the databases are surrounded by a CGI wrapper, which controls what data are accessed for a particular query initiated through a web form.

This architecture is based on necessity rather than design. When the RCSB took over the PDB the limited time available for providing a working system necessitated the joining together of components that we had used in our individual laboratories. This has served us well, but represents issues in maintenance, portability and extendibility. A completely re-engineered system is in an alpha stage of development and consists of a three-tier system using Enterprise Java. This system will go into beta testing in 2004. The back-end (first tier) is a single relational database, which for production will use IBM DB2. Other databases may be easily substituted. For example, we also

A database resource is only as good as the data it contains regardless of the sophistication of the technology used to support it

have a version of the re-engineered site that uses MySQL. This free software will enable users to support a version of the PDB in their laboratories at no software cost. The middle tier is the web application, which is Java 2 Enterprise Edition (J2EE) compliant. The Java code, Java Server Pages, servlets, etc., run in a J2EE-compliant web server (currently JBoss version 3.2). Both web page requests (http) and web services requests (XML) are processed and answered by this tier. To facilitate the functions of this middle tier, such as searching, browsing and display, several open-source libraries are used:

- The code developed by the PDB for the middle layer is object-oriented Java, so the Hibernate library is used as an object-to-relational mapping package. This allows code in the middle tier to deal with macromolecular data in an object-oriented way, regardless of how the data are stored on disk. This decoupling allows for easier migration to other persistence options, for example MySQL introduced above or an XML database such as Apache Xindice.
- To provide Google-like response times for searching, the Lucene library is used for keyword searching. Appropriate fields are indexed with this package and the resulting index files are used in the keyword search portion of the web application. The response time, disk use and CPU use are much better than a SQL-based keyword search, and it offers a rich query syntax, for example, fuzzy searches, proximity searches, ranges searches, compound searches, term boosting and grouping.
- The Jakarta Struts library is used to implement a model–view–controller framework for the PDB website. Using Struts, all web pages are assembled dynamically using templates. This accommodates future website personalisation and makes website maintenance and modification easier.

A database resource is only as good as the data it contains regardless of the sophistication of the technology used to support it. This brings us to the most important activity undertaken by the PDB, where the mandate is not only to provide data in a timely way, but to provide data that are accurate, consistent and complete. Achieving this is what we refer to as data uniformity and has consumed approximately one-third of PDB resources over the past five years and is only now becoming apparent to users.

THE DATA

The lack of uniformity in PDB data is a consequence of a long history of collection during a period when the fields of structural biology and biology have changed rapidly

The lack of uniformity in PDB data is a consequence of a long history of collection during a period when the fields of structural biology and biology have changed rapidly. The original archivists could not have conceived of the types of comparative analysis that would be demanded of these data, nor did they have the information technology available to record the data as consistently as is needed today. The exchange dictionary and associated tools permit us to provide a consistent record of the data collected today, but still requires that we remediate the data collected before the availability of this dictionary to conform to the current standard. Part of this process is described elsewhere[30,31] and only an indication of what is required is given here. Consider as an example the use of the Enzyme Classification. Initially PDB entries either did not include these data, or the level of classification that exists today did not exist when the data were collected. To provide all enzyme structures with a current representation required that the PDB review each entry, often against annotation present in other databases such as Swiss-Prot, to define a current list. The mmCIF versions of these files and the re-engineered PDB database contain the results of this remediation that covers compound naming, resolution, source

organism and a number of other features. The experimental data from the X-ray experiment, namely the structure factors, have also been checked. Structure factors conform to a standard format and have been validated as far as possible.

The PDB maintains a historic archive

It is important also to preserve the original records associated with a structure deposition, as this is the original record that links the data to the published literature. Today this record, which includes interactions between the depositor and the PDB annotator, is electronic. Earlier work was on paper, magnetic tape and other media. The PDB maintains a historic archive of these materials, which are frequently called upon in the annotation of current entries to help provide a consistent record of current and past structures.

CONCLUSIONS

There are many facets to maintaining a biological data resource

What we have tried to convey is that there are many facets to maintaining a biological data resource for use by a diverse community of users. Only some are technical. There are sociological, managerial and political issues that also come into play. Key elements for success are good communications among those running the resource, who need to have diverse skill sets and among every member of the team and the communities they represent. Community feedback must be treated seriously and lead to a prioritised set of action items to be addressed by the resources available. The PDB uses a help desk, community listserver, focus groups and attendance at a variety of scientific meetings to solicit that feedback.

Beyond all else is the need for good data and a robust data representation

Beyond all else is the need for good data and a robust data representation that is flexible enough to meet the needs of a changing science. Together with appropriate use of current technologies, the resource should be able to respond to the needs of a changing community. At the PDB, we feel we will need to respond to many developments in the coming years, such as the growth of a more diverse user base that needs to have different and intuitive views of the PDB data and a different paradigm of access to PDB data. Today, users download PDB files and access them remotely from their own client applications. We see, through the emergence of web services and other technologies, a time when appropriate items of data are extracted as needed from the PDB and combined with data from other resources to give a current and integrated view of a biological macromolecule. For this vision to be realised we will need well-defined and published application programming interfaces (API) to the data and associated software to support such services. These are our obvious five year goals; there will be developments in our science and technology that are not so obvious. These make for an exciting future and something with which we feel confident can be dealt with given a foundation of solid data management practices.

Acknowledgments

The Protein Data Bank (PDB) is managed by three members of the Research Collaboratory for Structural Bioinformatics – Rutgers University, SDSC/UCSD and CARB/NIST – and is funded by the National Science Foundation, the Department of Energy and the National Institutes of Health.

The mmCIF working group members include Paula M. Fitzgerald, Helen M. Berman, Philip E. Bourne, Brian McMahon, Keith Watenpaugh, John Westbrook, Enrique Abola, Eleanor Dodson, Lynn Ten Eyck, Art Olson and Wolfgang Steigemann.

References

1. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.*, Vol. 28, pp. 235–242.
2. URL: http://www.pdb.org
3. Cold Spring Laboratory Press (1972), 'Cold Spring Harbor Symposia on Quantitative Biology', Vol. 36, Cold Spring Laboratory Press, Cold Spring Harbor.
4. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B. *et al.* (1977), 'Protein Data Bank: A computer-based archival file for macromolecular structures', *J. Mol. Biol.*, Vol. 112, pp. 535–542.
5. Kendrew, J. C., Bodo, D., Dintzis, H. M. *et al.* (1958), 'A three-dimensional model of the

myoglobin molecule obtained by X-ray analysis', *Nature*, Vol. 181, pp. 662–666.

6. Watson, H. C. (1969), 'The stereochemistry of the protein myoglobin', *Prog. Stereochem.*, Vol. 4, p. 299.
7. Perutz, M. F., Rossmann, M. G., Cullis, A. F. *et al.* (1960), 'Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution', *Nature*, Vol. 185, pp. 416–422.
8. Dodson, E., Harding, M. M., Hodgkin, D. C. and Rossmann, M. G. (1966), 'The crystal structure of insulin. 3. Evidence for a 2-fold axis in rhombohedral zinc insulin', *J. Mol. Biol.*, Vol. 16(1), pp. 227–241.
9. Deisenhofer, J., Epp, O., Sinning, I. and Michel, H. (1995), 'Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*', *J. Mol. Biol.*, Vol. 246(3), pp. 429–457.
10. Lesk, A. M. and Chothia, C. (1980), 'How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins', *J. Mol. Biol.*, Vol. 136(3), pp. 225–270.
11. Bourne, P. E. and Weissig, H., Eds (2003), 'Structural Bioinformatics', John Wiley & Sons, Inc., Hoboken, NJ.
12. Markley, J. L., Ulrich, E. L., Westler, W. M. and Volkman, B. F. (2003), 'Macromolecular structure determination by NMR spectroscopy', in Bourne, P. E. and Weissig, H., Eds, 'Structural Bioinformatics', John Wiley & Sons, Inc., Hoboken, NJ, pp. 89–113.
13. *Nature Structural Biology* (2000), 'Structural Genomics Supplement Issue', *Nature Struct. Biol.*, Vol. 7 (URL: http://structbio.nature.com/).
14. URL: http://www.wwpdb.org
15. URL: http://www.ebi.ac.uk/msd
16. URL: http://www.pdbj.org
17. Berman, H., Henrick, K. and Nakamura, H. (2003), 'Announcing the worldwide Protein Data Bank', *Nature Struct. Biol.*, Vol. 12, pp. 980.
18. International Union of Crystallography (1989), 'Commission on Biological Macromolecules', *Acta Crystallogr. A*, Vol. 45, Part A, p. 658.
19. Hall, S. R., Allan, A. H. and Brown, I. D. (1991), 'The Crystallographic Information File (CIF): A new standard archive file for crystallography', *Acta Crystallogr.*, Vol. A47, pp. 655–685.
20. Fitzgerald, P. M. D., Berman, H. M., Bourne, P. E. and Watenpaugh, K. (1993), 'The macromolecular CIF dictionary', *ACA Annual Meeting*, Vol. 21, Albuquerque, p. D008.
21. URL: http://deposit.pdb.org/mmcif
22. Westbrook, J. and Fitzgerald, P. M. D. (2003), 'The PDB format, mmCIF formats and other data formats', in Bourne, P. E. and Weissig, H., Eds, 'Structural Bioinformatics', John Wiley & Sons, Inc., Hoboken, NJ, pp. 161–179.
23. Bourne, P. E., Berman, H. M., McMahon, B. *et al.* (1997), 'The macromolecular Crystallographic Information File (mmCIF)', *Methods Enzymol.*, Vol. 277, pp. 571–590.
24. Hall, S. R. (1991), 'The STAR File: A new format for electronic data transfer and archiving', *J. Chem. Inf. Comput. Sci.*, Vol. 31, pp. 326–331.
25. Greer, D. S., Westbrook, J. and Bourne, P. E. (2002), 'An ontology driven architecture for derived representations of macromolecular structure', *Bioinformatics*, Vol. 18, pp. 1280–1281.
26. The Gene Ontology Consortium (2000), 'Gene Ontology: Tool for the unification of biology', *Nature Genetics*, Vol. 25, pp. 25–29.
27. Software developed by the PDB for use with mmCIF data files is available from http://deposit.pdb.org/software/.
28. Shindyalov, I. N. and Bourne, P. E. (1997), 'Protein data representation and query using optimized data decomposition', *CABIOS*, Vol. 13, pp. 487–496.
29. Apache Software Foundation (1999–2003), Lucene.
30. Bhat, T. N., Bourne, P. E., Feng, Z. *et al.* (2001), 'The PDB data uniformity project', *Nucleic Acids Res.*, Vol. 29(1), pp. 214–218.
31. Westbrook, J., Feng, Z., Jani, S. *et al.* (2002), 'The Protein Data Bank: Unifying the archive', *Nucleic Acids Res.*, Vol. 30, pp. 245–248.