WILEY InterScience® DISCOVER SOMETHING GREAT CHIRALITY 18:426–429 (2006)

Protein Circular Dichroism Data Bank (PCDDB): Data Bank and Website Design

LEE WHITMORE,[1] ROBERT W. JANES,[2] AND B. A. WALLACE[1,3]*

[1]*Dept. of Crystallography, Birkbeck College, University of London, London, U.K.*
[2]*School of Biological and Chemical Sciences, Queen Mary, University of London, London, U.K.*
[3]*Centre for Protein and Membrane Structure and Dynamics, Daresbury Laboratory, Warrington, U.K.*

Presented at the 10th International Conference on Circular Dichroism, 2005, Sandestin, Florida

ABSTRACT The Protein Circular Dichroism Data Bank (PCDDB) is a new deposition data bank for validated circular dichroism spectra of biomacromolecules. Its aim is to be a resource for the structural biology and bioinformatics communities, providing open access and archiving facilities for circular dichroism and synchrotron radiation circular dichroism spectra. It is named in parallel with the Protein Data Bank (PDB), a long-existing valuable reference data bank for protein crystal and NMR structures. In this article, we discuss the design of the data bank structure and the deposition website located at http://pcddb.cryst.bbk.ac.uk. Our aim is to produce a flexible and comprehensive archive, which enables user-friendly spectral deposition and searching. In the case of a protein whose crystal structure and sequence are known, the PCDDB entry will be linked to the appropriate PDB and sequence data bank files, respectively. It is anticipated that the PCDDB will provide a readily accessible biophysical catalogue of information on folded proteins that may be of value in structural genomics programs, for quality control and archiving in industrial and academic labs, as a resource for programs developing spectroscopic structural analysis methods, and in bioinformatics studies. *Chirality 18:426–429, 2006.*

KEY WORDS: bioinformatics; circular dichroism spectroscopy; synchrotron radiation circular dichroism (SRCD) spectroscopy; database; structural genomics; archive; quality control

The technique of circular dichroism (CD) has become an important tool to support, verify, and sometimes uniquely demonstrate many aspects of protein structure and function. At present, many structural biology publications include CD spectra, but often the data from such studies are not easily accessible to anyone wishing to use them. One way to tackle this problem is the creation of a central deposition facility where CD spectral data can be archived and readily and openly accessed. The Protein Circular Dichroism Data Bank (PCDDB) has been designed to be such a data bank resource for the deposition of both CD and synchrotron radiation circular dichroism (SRCD) spectral data, together with accompanying experimental information.[1]

In structural biology, once the field of protein crystallography had been developed, the Protein Data Bank (PDB)[2] was established as a deposition data bank for protein and nucleic acid structures. Later it was expanded to NMR structures, and the Biological Magnetic Resonance Data Bank (BMRB)[3] was created as a deposition data bank for macromolecular NMR structural data. Deposition of data in these data banks is now required by many journals at the time of publication. These data banks have proved to be rich resources for bioinformatics studies of various types. CD spectroscopy has also evolved as a critical focus of many structural biology studies; hence, the development of an equivalent data bank for this technique is particularly timely. In addition, as with the other data banks, the PCDDB will ultimately contribute to recent U.K. research council and U.S. funding agency[4] requirements for data archiving.

Since the collection of CD spectra tends to be a relatively simple process, this can sometimes lead to problems associated with some published data. A second facet of the PCDDB is to address this issue by enabling validation of spectra to be included in the data bank. Validation software will check the data to confirm that a set of criteria is met, thereby helping safeguard the quality of spectra incorporated into the data bank. It is expected that having validation standards will encourage the use of 'good practice' protocols in CD data collection. In essence, this aims to work in a parallel manner to the validation pro-

*Correspondence to: B. A. Wallace, Department of Crystallography, Birkbeck College, University of London, London WC1E 7HX, U.K.
E-mail: ubcg25a@mail.cryst.bbk.ac.uk or to:
R. W. Janes, School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, U.K.
E-mail: r.w.janes@qmul.ac.uk
Received for publication 16 December 2005; Accepted 16 January 2006
DOI: 10.1002/chir.20267
Published online 12 April 2006 in Wiley InterScience (www.interscience.wiley.com).

cedures introduced into X-ray and NMR structure determination through packages such as PROCHECK[5] and WHATIF,[6] which significantly contributed to raising the quality of structures deposited in the PDB. The PCDDB validation software will also be valuable as standalone software for quality control and cross-calibration purposes.[7]

In parallel with the PDB, spectral data from proteins, nucleic acids, peptides, and (uniquely here) sugars will be included in the PCDDB data bank. It is anticipated that the PCDDB will open up a rich vein of data-mining and bioinformatics research, offering new insights into the relationships between structure and function so fundamental to today's structural biology community. For example, with an abundance of CD and SRCD spectra in the PCDDB, many protein structural folds will be represented and this could provide a ready way of detecting new folds and other structural features, especially from SRCD spectra, which include the data at lower wavelengths[8] associated with charge transfer transitions. Another important application for the PCDDB data bank will be as a repository for data from structural genomics programs. As these programs aim to clone and express large numbers of new soluble and membrane proteins, and as CD (and especially SRCD) requires only very small amounts of protein,[9] it will be possible to facilely document spectra of all expressed proteins and then ultimately link the PCDDB files to PDB files, once their crystal or NMR structures have been determined.

Of major importance in creating such a resource, therefore, is how the data bank is constructed and accessed, and what its contents will be. A detailed list of proposed information to be included in each entry has been published,[1] including modifications which arose from discussions with the PCDDB International Scientific Advisory Board[10] and additional comments provided through an open consultation with the structural biology community. It is expected that the contents will constantly evolve and will thus require a flexible database structure to adapt to further modifications.

This article addresses the structure of the data bank and website. The development is now at the stage of creating a prototype and initial working model. The design considerations for the data bank and website outlined in this article are open to public comments at pcddb@mail.cryst.bbk.ac.uk.

DATA BANK DESIGN

The pilot stage of the data bank design has been driven by functionality and achievability, with computational performance a lesser, but not insignificant, consideration. During the pilot phase, this approach and the limited volume of data to be input from alpha testing sites will allow for rapid development, while not significantly affecting the data bank performance. Computational optimization of the underlying database architecture will, however, be part of the later large-scale development work.

One of the features of a data bank not found in a database is the ability to allow authorised users to deposit data into the data bank. This functionality requires more computational procedures than those of a simple database, including such tasks as user account creation, deposition cross-checking, and data validation. These procedures generally incorporate a data curation aspect, which can be performed manually, or semi- or fully-automatically. The design of the PCDDB extra-data bank procedures has been focused on automatic curation for as many aspects as possible. The rationale for this is that in the long term the development time devoted to enabling computational curation will ultimately be far more efficient than the time that would be expended in manual curation on each entry, even though the latter would be expected to be less demanding in the pilot phase. Furthermore, although the data bank is being developed at a single site, it is planned that continuing and long-term curation, user deposition, archiving, and data backup will be at a number of distributed sites, most notably at SRCD facilities around the world.[10] As a result, procedures must eventually minimize manual curation and facilitate coordinated curation at multiple sites; hence, the importance of automated curation facilities.

Data in the data bank will be held in one of three states: validated, nonvalidated, and 'record-metadata.' The validated data will be publicly accessible and non-amendable and will fulfill the regulatory targets that the data bank aims to serve. Nonvalidated data will be accessible only to the depositor of the data and will be amendable, as this will represent data at various stages of being processed. This state of data will allow for functionality such as partially completed submissions and delayed-release-date submissions. Record-metadata will adopt the same accessibility status of the record to which it is attached and will be amendable both by the depositor of that data (noted by dated comment) and the data bank curators. The purpose of this record-metadata is to store information that may change over time, despite the validated record itself being nonamendable. Examples of such information would be to indicate if a record has been made scientifically obsolete by deposition of a newer record (as is the case in the PDB) or to add new links to information about the record that may subsequently appear in other facilities or data banks. Record-metadata is itself distinct from the metadata produced in the circular dichroism experiments, which in the context of the data bank is part of the validated data and is nonamendable experimental metadata.

Individual records will be constructed with a majority of data being in the validated or nonvalidated state, accompanied by a small portion of record-metadata. The vast majority of records in the data bank will be of the validated data plus record-metadata type, with only a small proportion existing at any particular time in the nonvalidated data plus record-metadata format. This latter type of record would be expected, in due course, either to be converted into validated records, or (possibly) deleted by either the depositor or (potentially) the data bank curators. Deletion of nonvalidated records would take place either at the request of the depositor or upon the closure of the depositor's PCDDB account after a predefined time period for completion had expired.

Data in the individual records will be held under nine grouped classifications,[1] with each classification bringing together related aspects of the record. The classifications are: sample, experimental conditions, spectra, calibration, data processing procedures, secondary structure, metadata, record-metadata, and depositor/citations. The contents of the classifications as they currently stand are slightly modified from the previously published list[1] and are noted on the prototype website at http://pcddb.cryst.bbk.ac.uk/. They are subject to revision as the data bank is developed; therefore, an important feature will be flexibility in the data bank structure so as to enable ease of modification at a later date.

WEBSITE DESIGN

The website for the PCDDB is the primary user interface for the data bank. In the pilot phase, there is no intention of providing any other interface methods (such as distributing the data via compact disk, DVD, or by binary file transfer), although these are possibilities in the later working stages of the development. To this end, the website aims to provide all of the deposition and searching functionality of the data bank. The primary concerns of the website are ease of use and functionality.

A user-account system has been devised for the PCDDB that will be mandatory for deposition but optional for searching. For users wishing to make depositions, a PCDDB account will allow storage and retrieval of partially submitted records as well as automated communications about the status and progress of submissions. For researchers whose primary interest might be in searching the data bank, creating a PCDDB user account would facilitate stored searches, automated searches, alerts, and opt-in informational newsletters, and thus generally enhance the experience of using the PCDDB. The casual user would be able to search the data bank through the website without setting up a user account, but would not have access to the additional support facilities as a result.

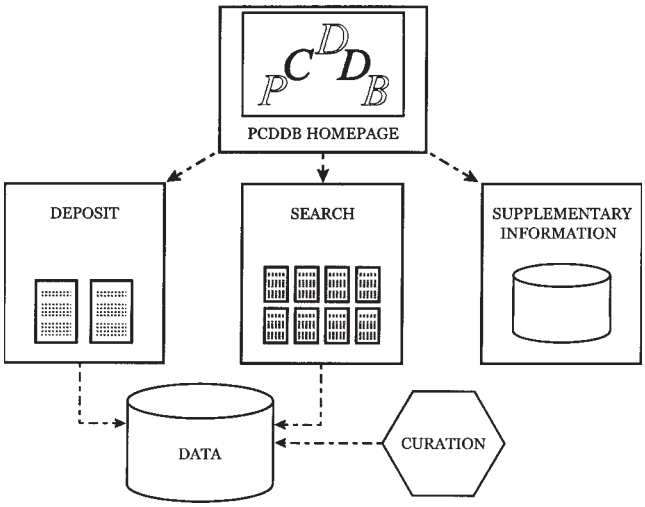

Fig. 1. Schematic overview diagram of the PCDDB website layout.

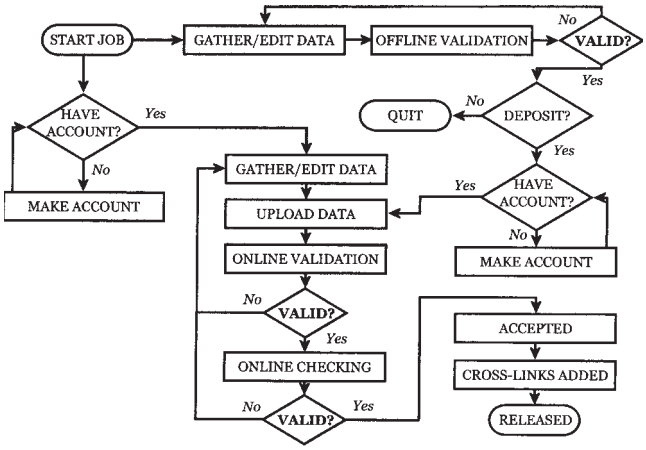

Fig. 2. Flow diagram of data deposition and validation procedures for the PCDDB.

The overall design of the website is focused on aiding the usability of the PCDDB. To this end, the design strategy of the site is to have three clearly defined areas: deposition, search, and supplementary information. These simple logical divisions neatly divide the site into its component sections and should provide an easy-to-use end product. Figure 1 shows an overall schematic representation of the website layout and procedures.

The deposition section of the website features the data-deposition interface, the data-validation tools (being developed as both online and downloadable, standalone versions), and an interface for storing and retrieving partially completed depositions. Figure 2 shows how data will ultimately reach and be incorporated into the data bank.

The search section of the website features a set of predefined commonly used search patterns such as search by protein name, author name, keyword, deposition date, or identifier. These searches will be simple to use and require minimal knowledge of the structure of the data bank. More specific searches can be tailored for users' needs in the advanced search feature, and will enable searches by, for example, secondary structure content, sequence, wavelength range, wavelength minimum and maximum, and other parameters that may be of use for bioinformatics and theoretical studies. Also, to facilitate data-mining and customised queries in the searching section, an SQL query feature will be implemented. A guide to the underlying construction of the data bank will allow users to perform more detailed searches of the records within the data bank via their own scripts (written in languages such as Perl or Java). All of the validated data and the record-metadata attached to validated records will be searchable.

The section of the website termed supplementary information will contain the remaining nondata information required by the site, such as user account creation and configuration interfaces, a site map, a glossary, and the obligatory but important website information such as terms and conditions, credits, and contact details. Information about CD spectroscopy, secondary structure calculations, and references, which is available on the

DICHROWEB calculation server,[11] will also be available in the supplementary section.

CONCLUSIONS

This article has described the development of a pilot version of the PCDDB, focusing on the considerations with regard to the data bank and website designs. These designs will be subjected to alpha-testing by a limited number of CD spectroscopic labs before being released publicly on the website located at: http://pcddb.cryst.bbk.ac.uk. It is anticipated that when it is fully operational, the website will become a valuable resource for the structural biology and bioinformatics communities. It will also become a deposition site for CD data generated by the various international structural genomics consortia as well as individual structural biology labs. Eventually, our aim is to expand the holdings in the data bank to include other data from vibrational CD, Raman optical activity, and Fourier transform infrared spectroscopy, for example, thus providing a wide-ranging archive for these techniques used to characterize biomacromolecules.

ACKNOWLEDGMENTS

This work was supported by a project grant from the BBSRC to B.A. Wallace and R. W. Janes, an International Workshop Grant from the BBSRC to B.A. Wallace, and an International Collaboration in Structural Genomics travel grant from the Foreign and Commonwealth Office to B.A. Wallace and R. W. Janes. We thank the members of the International Scientific Advisory Board of the PCDDB, Drs. F. Formaggio (Italy), K. Gekko (Japan), N. Greenfield (USA), S. Kelly (UK), J.-C. Maurizot (France), N. Price (UK), A. Rodger (UK), and J. Sutherland (USA); Drs. Helen Berman and John Westbrook of the Research Collaboratory in Structural Bioinformatics (PDB); and Dr. Hakuri Nakamura of the PDBj for helpful discussions.

LITERATURE CITED

1. Wallace BA, Whitmore L, Janes RW. The protein circular dichroism data bank (PCDDB): A bioinformatics and spectroscopic resource. Proteins: Struct, Funct Bioinf 2006;62:1–3.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
3. Seavey BR, Farr EA, Westler WM, Markley JL. A relational database for sequence-specific protein NMR data. J Biomolecular NMR 1991; 1:217–236.
4. NIH Notice NOT-OD-03–032. Sharing research data. 2003.
5. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK - A program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
6. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graphics 1990;8:52–56.
7. Guideline Q6B, International Conference on Harmonisation of Techincal Requirements for Registration of Pharmaceuticals for Human Use; 1999. Brussels. FDA Register 64FR. p. 44928.
8. Wallace BA, Janes RW. Synchrotron radiation circular dichroism spectroscopy of proteins: Secondary structure, fold recognition, and structural genomics. Curr Opin Chem Biol 2001;5:567–571.
9. Wien F, Wallace BA. Calcium fluoride micro cells for synchrotron radiation circular dichroism spectroscopy. Appl Spectroscopy 2005; 59:1109–1113.
10. Wallace BA, Janes RW. International workshop on the protein circular dichroism data bank. Synchrotron Radiation News 2005;18:20–21.
11. Whitmore L, Wallace BA. DICHROWEB: an online server for protein secondary structure analyses from circular dichroism spectroscopic data. Nucleic Acids Res 2004;32:W668–673.