Structure
Ways & Means

The Protein Circular Dichroism Data Bank, A Web-Based Site for Access to Circular Dichroism Spectroscopic Data

Lee Whitmore,[1] Benjamin Woollett,[1,3] Andrew J. Miles,[1,3] Robert W. Janes,[2,*] and B.A. Wallace[1,*]
[1]Department of Crystallography, Institute of Structural and Molecular Biology, Birkbeck College, University of London, London WC1E 7HX, UK
[2]School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, UK
[3]These authors contributed equally to this work
*Correspondence: b.wallace@mail.cryst.bbk.ac.uk (B.A.W.), r.w.janes@qmul.ac.uk (R.W.J.)
DOI 10.1016/j.str.2010.08.008

SUMMARY

The Protein Circular Dichroism Data Bank (PCDDB) is a newly released resource for structural biology. It is a web-accessible (http://pcddb.cryst.bbk.ac.uk) data bank for circular dichroism (CD) and synchrotron radiation circular dichroism (SRCD) spectra and their associated experimental and secondary metadata, with links to protein sequence and structure data banks. It is designed to provide a public repository for CD spectroscopic data on macromolecules, to parallel the Protein Data Bank (PDB) for crystallographic, electron microscopic, and nuclear magnetic resonance spectroscopic data. Similarly to the PDB, it includes validation checking procedures to ensure good practice and the integrity of the deposited data. This paper reports on the first public release of the PCDDB, which provides access to spectral data that comprise standard reference datasets.

INTRODUCTION

Circular dichroism (CD) spectroscopy is a widely used technique in structural biology for the characterization of protein structure, dynamics, and folding. To date, there has been no public means of access to any published (or unpublished) CD spectral data. The Protein Circular Dichroism Data Bank (PCDDB) (http://pcddb.cryst.bbk.ac.uk) has been designed as a web-based resource for accessing, storing, and depositing protein CD and synchrotron radiation CD (SRCD) spectra. It has been established to provide a central access point for CD and SRCD data in a manner parallel to that of the Protein Data Bank (PDB) (Berman et al., 2000, 2003, 2007) for protein crystal, nuclear magnetic resonance, and cryo-electron microscopy structural data.

The concept and initial design of the PCDDB were first described in 2006 (Wallace et al., 2006; Whitmore et al., 2006); the intervening period since then has been used for extensive open and public consultation regarding the content, format, and validation procedures. This paper details the first public release of the project. It describes the means of accession, features, functionalities, formats, and searching procedures, as well as the current contents of the data bank.

DISCUSSION

The PCDDB has been developed as a resource for the structural biology and bioinformatics communities. The records it contains will provide a readily accessible catalog of information on protein spectral characteristics that will provide standards for many types of experiments, including protein classification and identification studies, and for the development of new spectral methodologies, including investigations of CD parameters and ab initio calculations, as well as new empirical methods for secondary and tertiary structure analyses. As was the case with the PDB, its availability is also likely to lead to a number of other heretofore-unimagined applications, especially in the field of bioinformatics. This paper describes the first public release of the data bank.

EXPERIMENTAL PROCEDURES

For each deposited protein, the PCDDB entry includes both spectral data and associated metadata (Table 1) on sample identity, experimental conditions (including calibration details), sample contents, data processing procedures, and links to sequence (Apweiler et al., 2004) and structural (Berman et al., 2007) data bases (if available), as well as links to cited source publications. It also includes information on the secondary structure derived from the cognate PDB file (when available) calculated by using the defined secondary structure of protein (DSSP) algorithm (Kabsch and Sander, 1983), and, where available, links to secondary data bases containing the enzyme classification (E.C.) numbers (Bairoch, 2000) and CATH protein fold classification (Orengo et al., 1997). Each field has an associated information button describing its contents and attributes. A search facility is provided that operates on each of the metadata fields.

The spectral data are available both as the original raw data and as the final processed data. The associated high voltage (or dynode or high tension) spectrum, which is effectively a pseudo-absorbance spectrum collected at the same time as the CD data, and which is important for monitoring the quality and validity of the data, is also available. Each record also includes the relevant instrument calibration spectral files. Users can view and download the spectra as a printable graphics file (in .gif format) and/or download the spectral data and the metadata in generic text format (with the suffix .pcd). The format of the file is indicated in an easily recognizable, variable-length header section, containing key-value pairs, split after 60 characters, followed by a multicolumn data section. Alternatively, users can download the spectral and header data in the .gen format, used by the CDtools processing and display software (Lees et al., 2006). Users also have the option to download the entire contents of the data bank as a single compressed archive.

Table 1. Experimental Parameters and Other Metadata Associated with Each PCDDB Entry

Data Type				
Sample Contents	Experimental Parameters	Instrument Calibration Data	Data Processing	Protein Information
Protein name[a]	Instrument[a]	Calibration compound(s)	Molecular weight[a]	PDB ID (if any)[a]
Source organism[a]	Protein concentration[a]	Concentration	Number of residues[a]	Uniprot ID ([a] or sequence)
Expression system	Protein purity	Cell pathlength	Mean residue weight[a]	Sequence ([a] or Uniprot ID)
Construct/modifications present	Buffer contents[a]	Temperature	Software name and version[a]	E.C. number
Ligands present	Baseline contents[a]	Date measured	Smoothing details (if done)[a]	CATH code
Macromolecular partners present	Temperature[a]	CSA ratio	Zeroing details (if done)[a]	Medline entry (when available)[a]
	Detector angle		Units[a]	Keywords
	Sample cell type[a]			Publication details[a]
	Sample cell pathlength[a]			
	Number of repeat scans[a]			
	High and low wavelengths[a]			
	Wavelength interval[a]			
	Instrument scan parameters[a]			
	Collection date[a]			

[a] Required items.

All spectra held within the data bank are subject to a validation procedure to ascertain that entries are of high quality and to ensure the integrity of the data bank. This validation procedure has analogous aims to those of the WHAT_CHECK (Hooft et al., 1996), MolProbity (Davis et al., 2004), and PROCHECK (Laskowski et al., 1993) validation procedures used for checking protein crystal structure determinations, the reports of which are linked to the PDB record for that structure. The PCDDB checks include completeness of the entry, spectral quality, suitability of the experimental measurements, calibration standards, and the values of the parameters used in the calculations, and are based on best practice considerations established for CD (Jones et al., 2004; Kelly et al., 2005) and SRCD spectroscopy (Miles and Wallace, 2006). A validation report is an integral part of the standard PCDDB record, and is included in the Web interface as the final data table. The Website tool-tips associated with each validation criterion in the table describe what feature has been tested and the expected values. In order to record accurately the validation criteria used against each record in the PCDDB, the validation report linked to the deposition record also contains a date and version stamp.

The initial data deposited into the PCDDB are SRCD spectra from the 71 soluble proteins that comprise the reference dataset collectively known as SP175 (Lees et al., 2006), which is now widely used for CD- and SRCD-derived secondary structure analyses (Whitmore and Wallace, 2006). These spectra were selected because they represent a sizable collection of high-quality spectra derived from proteins that have corresponding high-quality crystal structures present in the PDB. Because the spectra were collected with SRCD beamlines, they all include low-wavelength data not normally achieved with conventional CD instruments, but all have been cross-correlated with standard CD instruments, and are indistinguishable from conventional CD spectra in the far-ultraviolet wavelength region (Lees et al., 2006). The reference dataset derived from these spectra (designated SP175) is currently available as a reference set in the DichroWeb CD analysis deconvolution server (Whitmore and Wallace, 2004). Phase one public access to the PCDDB now makes the individual spectra and their experimental conditions (including the raw data) publicly accessible, downloadable, and searchable.

Phase two public release of the PCDDB with a public deposition interface is planned for early 2011 (after feedback from users on the accessibility, functionality, and contents available in phase one). It will enable users to deposit their spectra into the data bank, which will augment significantly the holdings of the PCDDB. The PCDDB will be a permanent repository for spectra, and will serve as a facility for data sharing (and, thus, be a simple means of fulfilling international granting body requirements) and data mining, and allow a wide range of new applications in bioinformatics and structural biology. Future developments will include computational tools for spectral matching, back calculations of spectra from crystal PDB files, enhancements to the validation procedures, and the ability to compile user-defined reference sets from the PCDDB, which may be used in DichroWeb and other deconvolution packages. A variety of other analytical tools, including cluster analysis techniques currently available in the CDTools software package (Lees et al., 2004), will be made available at the PCDDB Website. There will also be tools for sorting of entries based on secondary structure, protein sequence homology, and spectral characteristics. Users will also be able to conduct statistical analyses on the entries within the PCDDB, and it is anticipated that this resource, together with the other analyses packages, will greatly enhance the information available from CD and SRCD data.

ACKNOWLEDGMENTS

We thank Daniel Klose of Queen Mary, University of London for helpful suggestions and for checking the DSSP calculations for the entries. This work was supported by the U.K. Biotechnology and Biological Sciences Research Council (grants no. F010346 and F010342 to B.A.W. and R.W.J., respectively, from the Bioinformatics and Biological Resources Fund, and grant no. G023476 to B.A.W. from the Tools and Resources Development Funding stream). The authors declare they have no conflict of interest.

Received: May 27, 2010
Revised: August 2, 2010
Accepted: August 13, 2010
Published: October 12, 2010

REFERENCES

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H.Z., Lopez, R., Magrane, M., et al. (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *32*, D115–D119.

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res. *28*, 304–305.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. *35*, D301–D303.

Davis, I.W., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2004). MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res. *32*, W615–W619.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381*, 272.

Jones, C., Schiffmann, D., Knight, A., and Windsor, S. (2004). Val-CiD best practice guide: CD spectroscopy for the quality control of biopharmaceuticals. National Physical Lab Report DQL-AS 008 (Teddington, UK: National Physical Laboratory).

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kelly, S.M., Jess, T.J., and Price, N.C. (2005). How to study proteins by circular dichroism. Biochim. Biophys. Acta *1751*, 119–139.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291.

Lees, J.G., Smith, B.R., Wien, F., Miles, A.J., and Wallace, B.A. (2004). CDtool—an integrated software package for circular dichroism spectroscopic data processing, analysis and archiving. Anal. Biochem. *332*, 285–289.

Lees, J.G., Miles, A.J., Wien, F., and Wallace, B.A. (2006). A reference database for circular dichroism spectroscopy covering fold and secondary structure space. Bioinformatics *22*, 1955–1962.

Miles, A.J., and Wallace, B.A. (2006). Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. Chem. Soc. Rev. *35*, 39–51.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997). CATH—a hierarchic classification of protein domain structures. Structure *5*, 1093–1108.

Wallace, B.A., Whitmore, L., and Janes, R.W. (2006). The Protein Circular Dichroism Data Bank (PCDDB): a bioinformatics and spectroscopic resource. Proteins *62*, 1–3.

Whitmore, L., and Wallace, B.A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. Nucleic Acids Res. *32*, W668–W673.

Whitmore, L., and Wallace, B.A. (2006). Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. Biopolymers *89*, 392–400.

Whitmore, L., Janes, R.W., and Wallace, B.A. (2006). Protein Circular Dichroism Data Bank (PCDDB): data bank and website design. Chirality *18*, 426–429.