SHORT COMMUNICATION

The Protein Circular Dichroism Data Bank (PCDDB): A Bioinformatics and Spectroscopic Resource

B. A. Wallace,[1,2]* Lee Whitmore,[1] and Robert W. Janes[3]*

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

ABSTRACT: **This article describes the development and creation of the Protein Circular Dichroism Data Bank (PCDDB), a deposition and searchable data bank for validated circular dichroism spectra located at http://pcddb.cryst.bbk.ac.uk/. Proteins 2006;62:1–3.** © **2005 Wiley-Liss, Inc.**

Key words: data bank; bioinformatics; circular dichroism spectroscopy; Synchrotron Radiation Circular Dichroism (SRCD); protein structure; Protein Data Bank (PDB)

INTRODUCTION

Circular Dichroism (CD) spectroscopy is an established and valuable technique for examining protein structure, dynamics and folding and is regularly used as a standard method in a large number of biological and chemical laboratories; new applications for this technique[1] are emerging as a result of the development of synchrotron radiation circular dichroism (SRCD) instrumentation in recent years.

At present there is no central resource or means of public access to published CD data files. We report here on the development and creation of the Protein Circular Dichroism Data Bank (PCDDB), a deposition and searchable data bank for validated circular dichroism spectra of biomacromolecules, located at http://pcddb.cryst.bbk.ac.uk, which should be a useful resource for structural molecular biology. Its aim is to provide open access and archiving facilities for circular dichroism spectra, in parallel to the Protein Data Bank (PDB), a long-existing and valuable reference data bank for protein crystal and NMR data.[2] It is anticipated that the PCDDB will become a valuable resource and of significant benefit to both the spectroscopic and wider structural biology and bioinformatics communities.

A prototype is currently accessible at http://pcddb.cryst.bbk.ac.uk, and detailed lists of proposed contents and validation parameters are included on the website and in Tables I and II of this Correspondence. The current version has incorporated advice from the members of the PCDDB International Scientific Advisory Board. An open consultation on the contents, validation procedures, and access will be available by email to pcddb@mail.cryst.bbk.ac.uk for a period of two months following the publication of this communication.

The process of spectral deposition to the data bank is designed to be via a user-friendly web site and in the future will include automatic reading of a range of data formats and data mining from file headers[3] to facilitate the process. Entries will be linked, in the case of spectra of proteins whose structures and sequences are known, to the appropriate PDB[2] and sequence data bank files.[4]

The PCDDB data bank entries include information on:

1. the protein (including links to sequence and structure data banks),
2. the sample (including methods and parameters for concentration determinations, assays of purity, amino acid composition),
3. spectral conditions and parameters,
4. instrument calibration,
5. spectral processing procedures,
6. secondary structure analyses, and
7. references to the literature.

It incorporates net CD and HT (or dynode voltage) data in downloadable formats, and provides a formatted image of each of the spectra. A full listing of contents is included in Table I.

The PCDDB will include a series of validation tools and protocols (Table II) that provide reports on data quality (and will be accessible as stand-alone software). The included data must be accurately processed, standardized and validated in order to ensure integrity of the data bank

Grant sponsor: BBSRC project grant (to B.A.W. and R.W.J.) and International Workshop Grant (to B.A.W.)

*Correspondence to: B. A. Wallace, Department of Crystallography, Birkbeck College, University of London, London WC1E 7HX U.K. or R. W. Janes, School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 7NS U.K. E-mail: ubcg25a@mail.cryst.bbk.ac.uk (B.A.W) or r.w.janes@qmul.ac.uk (R.W.J.)

Received 4 May 2005; Accepted 14 June 2005

Published online 21 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20676

TABLE I. Contents and Parameters Included in the PCDDB†

A. Sample
- PCDDB identifier code
- Protein Name
- *Alternative name(s)*
- Protein CODE (Swiss-Prot, and where possible, PDB) [clickable links]
- Key Words (up to 10)
- Organism, Organ, Isoform
- Source (cloned, synthesized, isolated, commercial source)
- Wild Type/Mutant/Cloning Variants
- *Ligands Present (if any)*
- Depositor name and contact information
- Publication Reference: Authors, Journal, Date, Title, Pages [clickable link]

B. Experimental Details
- Protein Concentration and Quantitation Method
- *Purity (%, method of determination)*
- Buffer Contents
- Baseline Contents
- Temperature
- Sample Cell Pathlength
- Method used to calibrate Sample Cell Pathlength
- Sample Cell Type/Material
- Instrument/Model Number/ or SRCD Beamline
- *Local spectrum identifier*
- Date Collected
- Nitrogen Purge or Vacuum
- *Detector angle (if relevant)*
- Dwell Time/Scan Speed/Time Constant
- Wavelength Range (Max, Min)
- Wavelength Interval
- *Spectral Resolution*
- *Low Wavelength Cutoff*
- *Criterion for Low Wavelength Cutoff (HT or dynode value)*
- Number of Repeats
- CSA calibration:
  - Parameters: Concentration, Pathlength, Zero Point, Date Measured
  - Values: Ratio, Molar Ellipticity at 285 nm
- *Other Instrument Calibration Standards: [choice]*

C. Data Processing
- Data Processing Software/Version Number
- Smoothing—Yes/No—Number of Points/Algorithm Used
- Wavelength or Range Used for Zeroing
- MRW
- Units
- *Results: Secondary Structure Calculations [user or PCDDB provided]*
  - *Method Used*
  - *Calculation Software and Version Number*
  - *Reference Database Used*
  - *Wavelength Range Used*
  - *NRMSD or Other Goodness-of-fit Parameter*
  - *Percentages and Types*

D. Files
- CD spectrum (either processed *with error bars* and details of processing, or raw data with separate baseline file) [clickable link to image of spectrum]
- HT or Dynode Spectrum
- *CSA or Other Standard Sectrum*
- *Publication Reference .pdf file*

†Some parameters may be read from the file headers, some are optional (italics), and some will be provided by the PCDDB [ie. links]. Many parameters will be accessible from pull-down lists.

as a source of structural information for data mining. This type of quality control has, for the most part, been missing from CD data collection and publications to date. Again, this will parallel the development of crystallographic validation software such as WHATIF[5] and PROCHECK,[6] which have proved to be of considerable value not only for

TABLE II. Validation Parameters and Checking Tools in the PCDDB

Mean Residue Weight (MRW) value
$\Delta\varepsilon$ calculation
$\Delta\varepsilon$ values too large or too small
Standards (CSA/ACS) ratio values
Zeroing point ellipticity
Signal/noise too low
Baseline component mismatch
Smoothing too severe
HT or dynode limit exceeded
Secondary structure goodness-of-fit parameter too high

the deposited files but also for enhancing standards within the field.

It is anticipated that this data bank will provide a readily accessible biophysical catalogue of information on correctly folded proteins, for tracability, quality assurance, and archiving in industrial and academic labs, a data bank for investigations of CD parameters/*ab initio* calculations, a reference and deposition site for proteins examined as part of structural genomics programs, an accessible source of information on protein standards, a resource for programs developing spectroscopic secondary and tertiary structural analysis methods, and in a wide range of structural biology studies. As was the case with the PDB, after it comes into general use, it is likely to lead to a number of other heretofore unimagined applications, especially in the field of bioinformatics. It could also provide a ready means of fulfilling (UK) Research Council and US (NIH) public archiving requirements,[7] and provide a traceable resource for ICH Guidelines for Biological Products.[8]

In summary, the PCDDB and validation techniques described here have the potential to become important resources for the structural biology community. The validation software will enable "good practice" methodologies to be adopted throughout the CD data collecting community. The data bank should be a useful archive for CD data and enable bioinformatics mining of an, as yet, untapped source of data. The complementarity and links to other structure and sequence data banks should ensure that the PCDDB becomes a valuable component of structural genomics programs. Finally, although this is has been developed specifically for CD spectroscopy, it has the potential to ultimately be expanded to include other spectroscopic data such as Fourier transform infrared, Raman optical activity, and vibrational CD.

ACKNOWLEDGMENTS

We thank the members of the PCDDB International Scientific Advisory Board [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)] for their suggestions and input, and Drs. Helen Berman and John Westbrook of the RCSB for helpful discussions.

REFERENCES

1. 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.
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. Lees JG, Smith, BR, Wien F, Miles AJ, Wallace BA. *CDtool*—an integrated software package for circular dichroism spectroscopic data processing, analysis and archiving. Anal Biochem 2004;332: 285–289.
4. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
5. Vriend G. WHATIF—a molecular modeling and drug design program. J Mol Graphics 1990;8:52–56.
6. 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.
7. NIH Notice NOT-OD-03-032. Sharing research data. 2003.
8. Guideline Q6B, International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. 2001.