Beilstein-Institut

Experimental Standard Conditions of Enzyme Characterizations,
September 13th – 16th, 2009, Rüdesheim/Rhein, Germany

Standard Formats for Presentation of Spectroscopic Data on Enzymes

Richard Cammack

Pharmaceutical Sciences Research Division, King's College London,
150 Stamford Street, London SE1 9NH, U.K.

E-Mail: richard.cammack@kcl.ac.uk

Received: 10th February 2010 / Published: 14th September 2010

Abstract

Spectroscopic methods are often used to follow the course of enzyme-catalysed reactions. UV/visible spectrophotometry is the most common, but a wide range of other spectroscopic techniques, including infrared and nuclear magnetic resonance, as well as mass spectrometry, are in use. Spectroscopy and spectrometry are also used in the characterization of the enzymes themselves, and in the identification and quantification of substrates and cofactors. Hitherto, there has been no formal requirement to archive original spectroscopic data, as there is for protein structures and gene sequences. However, the funding agencies increasingly expect grantees to have policies on data sharing, and to deposit all types of experimental data. Spectroscopic data are now conveniently acquired in digital form, but apart from printed documents, there are no universally accepted formats for data storage. Spectra are produced by proprietary software written by instrument manufacturers to run their own instruments. Data formats are non-standard and may be difficult to read directly. Standard, vendor-neutral data formats have been established for certain types of spectroscopy, such as JCAMP-DX (from the Joint Committee on Atomic and Molecular Physical Data eXchange) for several types of spectroscopy, including infrared and NMR. The details of each format necessarily depend on the type of spectroscopy. There are parallel developments of criteria for meta-data and data validation. Standard formats will facilitate the use of electronic notebooks. The extension of these formats to different types of spectroscopy and spectrometry will facilitate their linkage to other chemical information such as molecular structure. ASCII formats such as JCAMP-DX or, more recently, XML formats such as CML (Chemical Markup Language) satisfy the data-storage requirements. They are readable by generic, open-source software. The routine deposition of spectra in electronic repositories (databanks) will benefit the biochemical community by making them available for further analysis and data mining.

Introduction

Many enzyme assays employ spectroscopy as a non-destructive method to measure kinetics of enzyme-catalysed reactions. UV/visible spectrophotometry is commonly used for continuous measurement of substrates or products. Fluorimetry offers higher sensitivity if fluorogenic substrates are available. For such purposes, the output is a two-dimensional graph of the concentration of a particular species as a function of time, from which reaction rates can be calculated. Often all that is required is a single absorbance reading at each time point. However modern spectrophotometers make it easy to measure the whole spectrum, which allows further operations such as spectral deconvolution and baseline subtraction to extract the concentration data from a sample containing multiple chromophores.

On early spectrophotometers, spectroscopic data were recorded on paper charts. The spectra were then transcribed into figures in paper publications. Operators became adept at recognizing the shapes and details of such spectra, but further analysis was very limited before spectrometers interfaced to computers were the norm. For publication of the data, the spectra were recorded at low resolution; they were often redrawn by hand, so that details were lost. The paper charts could be digitized, but this is laborious and entails loss of resolution. Meanwhile, the electronic versions of the spectra recorded on diverse computer systems soon were rendered unreadable by the rapid obsolescence of computer operating systems and data storage media. As a result, the data from many careful studies, sometimes on material that is no longer accessible, is now unavailable for analysis and comparison.

In the characterization of the enzymes and their cofactors, a wider range of spectroscopic methods is used [1], including NMR spectroscopy, Fourier-transform infrared (FTIR), circular dichroism (CD) and mass spectrometry. More specialist techniques can provide additional information about enzyme mechanisms, such as electron paramagnetic resonance (EPR) for flavins and transition-metal ions. Spectroscopic data are rich in information, and may be analyzed in different ways to extract information about enzymes and the reactions they catalyse, including:

- observation of transient intermediates in the enzyme-catalysed reaction;
- quantitative analysis, by comparison with spectra of standard samples;

- resolution of the spectra into their principal components, for example Gaussian line shapes;
- extraction of fundamental parameters of the enzyme-bound species, for example by simulation of the spectra using appropriate theory.

The Importance of Retaining Original Spectroscopic Data on Enzymes

During a typical study, many spectra are recorded. Techniques such as Fourier-transform NMR and pulsed EMR generate extensive sets of multidimensional data. Because of limitations of space, careful selection is required of data for publication, so the results are usually presented in the form of derived parameters. A whole spectrum may be reduced to a single data point in a two-dimensional plot. This was inevitable in the past, when the storage of large quantities of data was difficult and costly. Now that virtually infinite electronic storage capacity is available, the raw data can be preserved. Often acquired with much effort and expense, these are an extremely valuable resource for further analysis by different approaches. They are required for comparison with other experimental data, and further theoretical analysis.

From the point of view of the STRENDA initiative, it is important that when enzyme data that rely on spectroscopic techniques are published, full details of the experiment and results are provided. Ideally all the original spectral data should be made available, not just the derived data required to create the published figures [2, 3]. Journals, which require deposition of gene sequences or protein structures in open-access databases, do not mandate this for key spectroscopic data. They could not print all the experimental data from a study; for example, results that are valid but lead to negative conclusions are rarely published. In chemistry, it has been estimated that more than 99% of spectra that are recorded are lost [4]. Many data languish on computer disks “trapped by technical, legal and cultural barriers – a problem that open-data advocates are only just beginning to solve” [5]. However this situation is likely to change. Whereas in the past there has been little incentive to share such information, it is now becoming mandatory for publicly funded research to present all the experimental data. Many funding bodies now have policies that require data sharing.

This review considers the issues that have to be addressed for acquisition and long-term storage of spectroscopic data, with particular reference to those on enzymes. Seamless data transfer will be facilitated if spectrometers can output their data in standard formats; if the experimental data are recorded in electronic notebooks; and if repositories to store and retrieve the data are readily accessible on the internet. However the lack of data standards inhibits the full exploitation of spectroscopic data.

Digital Data Formats

Typically, the output from a spectrometer is captured as *x–y* data, such as a curve of absorbance *vs.* wavelength or frequency, as in UV/visible or FTIR spectrophotometry. In digital format this is stored in a series of data points representing a two-dimensional plot. Alternatively the output may take the form of a list of peak positions and amplitudes, as in NMR or mass spectrometry. Additional dimensions may also be introduced by varying other parameters, such as time or temperature, yielding larger data sets (*n*-tuple arrays). Other types of measurement, such as chromatography, also produce *x-y* plots. These data can be captured electronically and conveniently stored in computer databases and retrieved for further analysis. From these data, parameters are extracted for publication. Printed documents are a way to save the data in a permanent and easily readable format. However, if publications are the only source of spectroscopic results, most of the original information in the spectra is lost [2]. It is difficult to store and retrieve the original data, and in particular to search the spectra for particular features.

Digital data formats for spectroscopy were introduced by manufacturers of spectroscopic instruments, when these were interfaced to computers. This is very convenient; the instrumental parameters are stored automatically with the spectral data, and can be stored and exchanged between users of the same instrument. However incompatibility between the bespoke systems was an issue. When computer memory was at a premium, formats were highly compressed, such as binary data, and were not human readable. Moreover they were proprietary, needing expensive software to read them. The software written for one spectrometer is usually incompatible with data from other manufacturers' instruments, and even older instruments from the same manufacturer. This requires special programs to be written for inter-conversion of formats. Such software tends to become obsolete in time, as computer operating systems evolve.

Data Storage

After a time, retrieval of spectroscopic data raises other issues. The files need to be systematically documented. Whereas an expert human reader can readily recognize the subtle similarities and differences in shapes of different spectra, it is very difficult to program a computer to recognize the salient features automatically (famously, it is very difficult by means of computer software to distinguish a picture of a cat from a dog, the basis of the CAPTCHA security system [6]). Moreover, little information can be extracted from free-text fields. Therefore, meta-data are essential for the deposition and retrieval of the spectra. Moreover, there must be sufficient information for the work to be reproduced. For this reason, there has to be a formal system for deposition of information such as the details of sample, aims of the experiment, ownership of the data and credits for the work. This has to be entered by the operator, ideally at the time of the measurement.

Standard file formats

Sharing of spectroscopic data acquired on different instruments requires the introduction of a ‘‘lingua franca’’ or a standard, preferably non-proprietary, format. Otherwise it is necessary, for each manufacturer’s spectroscopic format, to have conversion software to and from other formats. In some areas of information technology such standard formats are well established, such as the portable document format (PDF) for printed documents, and the JPG and PNG formats for digital images. When such standards are established, files in these formats can be recognized by web browsers and other generic programs, which help to overcome the problem of obsolescence of specialist software.

Header	##TITLE = Nitrite reductase EPR ##JCAMP-DX = 4.24 $$ ##$URL =
Credits	##ORIGIN = King’s College London ##OWNER = Public Domain
Instrument	##DATA TYPE = EPR SPECTRUM
Measurement	##.FREQUENCY = 9.383 GHz ##NPOINTS = 1024 ##RESOLUTION = 3
Manipulation	##XFACTOR = 1.0 ##YFACTOR = 1.0
Spectrum	##XUNITS = mT ##YUNITS = ARBITRARY ##FIRSTX = 10 ##LASTX = 610 ##FIRSTY = -1.875 ##XYDATA = (X++(Y..Y)) 9.99 - 1.875 10.58 - 1.49 .. 608.45 2.585 609.03 2.725 ##END =
X – Y data	##XYDATA = (X++(Y..Y)) 9.99 - 1.875 10.58 - 1.49 .. 608.45 2.585 609.03 2.725 ##END =

Figure 1. Example of a spectrum, with the layout of the JCAMP-DX file.
The table shows some minimal information for presenting an EPR spectrum.

A number of different standard formats for spectroscopic data have been proposed over the years, with varying levels of acceptance by the community. JCAMP-DX is a flexible format, introduced in the 1980’s by the Joint Committee on Atomic and Molecular Physical Data.

It has been used extensively, notably for FTIR [7, 8], NMR [9] and mass spectrometry [10]. It consists of a single file, representing data arrays of 2, 3 or more dimensions. A file comprises labelled data records, each consisting of a flagged data label (usually fairly self-explanatory), and an associated data-set. The file may contain a spectrum or a block of spectra. It also incorporates meta-data to describe the sample(s), measurement conditions and other experimental details. The use of consistent data-labels makes it possible to use generic software to output from several different types of spectroscopy and spectrometry. An example, including some sections of a file for EPR, is shown in Table 1. Some spectrometer manufacturers, notably in FTIR and NMR, offer conversion software to produce spectra in JCAMP-DX format. The format is open-source, and can be read by any program that can handle ASCII text, though special software is needed to view and interpret the spectra. JSpecView is an Open Source applet for viewing JCAMP-DX and AnIML spectra files, allowing zooming and integration [11]. For studies of enzymes, JCAMP-DX protocols of interest in studies of enzymes are for infrared spectroscopy [7] (which can be extended to UV/visible spectrophotometry), NMR spectroscopy [9], electron magnetic resonance (EPR and ESR) spectroscopy [12] and mass spectrometry [10]

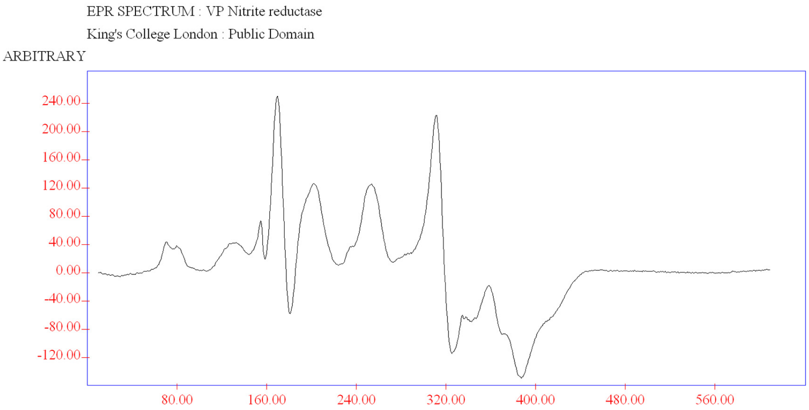

Figure 2. Spectrum plotted from the JCAMP-DX file in Figure 1. The table shows some minimal information for presenting a spectrum. It was plotted using the CHIME plug-in from Symyx [22] (note that this is no longer supported). Sample courtesy of Dr V.O. Popov, Bakh Institute of Biochemistry, Moscow.

Repositories for Spectroscopic data

Specialist technique-specific databanks are being developed for the purpose of collecting and disseminating spectroscopic data. Some of these databases are small enough that they can be downloaded from the internet. BioMagResBank – Biological Magnetic Resonance

Data Bank (BMRB: www.bmrb.wisc.edu) [13] – collects data on NMR spectroscopy of biological entities. It uses NMR-STAR format for proteins [14]. The Protein Circular Dichroism Data Bank (pcddb.cryst.bbk.ac.uk) is a newly-established repository for circular dichroism spectroscopy [15]. Ultraviolet CD is used for characterization of protein structure. It is used to estimate the proportion of secondary structure elements, such as alpha-helix and beta sheet. The PCCDB site is intended to offer an archive of user-deposited CD spectra; this data can be used in conjunction with analyses algorithms [16] to determine the secondary structures of proteins. It includes conversion software to convert spectra recorded in ten different formats on spectrometers from five manufacturers plus synchrotron CD spectra, into an in-house format that is downloadable in a generic text format.

XML formats

There are initiatives to replace the flat-file formats like JCAMP-DX by web-based markup languages. XML has been adopted in many areas of computer technology, including communications and distributed computing. There are XML schemas for various aspects of chemistry and biochemistry. The schemas incorporate data dictionaries to ensure consistent terminology, and validation criteria to ensure data quality. Chemical markup language, CML, is used to represent chemical data and documents [17, 18]. Within CML, an XML vocabulary, CMLSpect, has been developed for spectral data [2]. For analytical chemistry, AnIML (**An**alytical **I**nformation **M**arkup **L**anguage), first proposed in 2003, is under development as an IUPAC/ASTM unified standard for spectroscopic and related information [19, 20]. The five initial AnIML techniques are UV/visible and infrared spectrophotometry, mass spectrometry, 1D NMR and chromatography [21].

Once acquired, spectroscopic data may be stored in the institution where the work was done; with the publisher where the work was published; or in a central repository. In each case the repository should have facilities for data deposition, validation, viewing, search and retrieval. The SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data) Project, which uses markup languages [4] has pioneered the archiving of primary chemistry data, including crystallographic structures, spectroscopy (principally NMR) and computational chemistry. This initiative has identified a number of features that are necessary for such repositories. Each component has a unique and persistent identifier. There is an embargo system, for research data prior to publication. In addition to the metadata requirements already mentioned, there should be a graphical user interface (GUI) for browsing, navigation, and text-based searches of the text and metadata.

Conclusions

Deposition of spectroscopic data requires open-source, generic formats. These should be easy to read, archive and retrieve. Repositories must be able to accept multiple instrumental formats, including legacy data from older instruments. Software is required for archiving, retrieval, display and manipulation of spectroscopic data. Meta-data are essential for archiving and retrieval of data, and data-mining. To achieve this will require the expansion of deposition requirements for both published and unpublished data, and funding for software and database development.

Acknowledgements

I am greatly indebted to Robert Lancashire (University of the West Indies); Bonnie Wallace (Birkbeck College London) and Bob Janes (Queen Mary University London) and Peter Murray-Rust (University of Cambridge) for discussions and comment, and Dr Vladimir Popov for the data used in Figure 1.

Abbreviations

ASTM	American Society for Testing and Materials
BioMagResBank	Biological Magnetic Resonance Data Bank
CAPTCHA	Completely Automatic Public Turing Test to Tell Computers and Humans Apart
CD	circular dichroism
CML	chemical markup language
EMR	electron magnetic resonance
EPR	electron paramagnetic resonance
ESR	electron spin resonance
FTIR	Fourier-transform infrared spectroscopy
GUI	graphical user interface
IUPAC	International Union for Pure and Applied Chemistry
JCAMP-DX	Joint committee on atomic and molecular physical data exchange
MDL	MDL Molecular Design Ltd
NMR	nuclear magnetic resonance
PCDDB	Protein Circular Dichroism Data Bank
SPECTRa	Submission, Preservation and Exposure of Chemistry Teaching and Research Data
STRENDA	Standards for Reporting Enzymology Data

References

[1] Reymond, J. L. (2006) Enzyme assays: High-throughput screening, Genetic selection and Fingerprinting, Wiley-VCH, Weinheim.

[2] Kuhn, S., Helmus, T., Lancashire, R. J., Murray-Rust, P., Rzepa, H. S., Steinbeck, C., and Willighagen, E. L. (2007) Chemical markup, XML, and the world wide web. 7. CMLSpect, an XML vocabulary for spectral data. *Journal of Chemical Information and Modeling* **47**:2015 – 2034.
doi: http://dx.doi.org/10.1021/ci600531a.

[3] Cammack, R. (2010) EPR Spectra of Transition-Metal Proteins: the Benefits of Data Deposition in Standard Formats. *Applied Magnetic Resonance* **37**:257 – 266.
doi: http://dx.doi.org/10.1007/s00723-009-0095-2.

[4] Downing, J., Murray-Rust, P., Tonge, A. P., Morgan, P., Rzepa, H. S., Cotterill, F., Day, N., and Harvey, M. J. (2008) SPECTRa: The deposition and validation of primary chemistry research data in digital repositories. *Journal of Chemical Information and Modeling* **48**:1571 – 1581.
doi: http://dx.doi.org/10.1021/ci7004737.

[5] Nelson, B. (2009) Empty archives. *Nature* **461**:160 – 163.
doi: http://dx.doi.org/10.1038/461160a.

[6] Elson, J., Douceur, J. R., Howell, J., and Saul, J. (2007) Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. In *Proceedings of the 14th ACM Conference on Computer and Communication Security* (DiVimercati, S. D. C., Syverson, P., and Evans, D., eds) pp. 366 – 374, Alexandria, VA.

[7] McDonald, R. S., and Wilks, P. A. (1988) JCAMP-DX – a standard form for exchange of infrared-spectra in computer readable form. *Appl. Spectrosc.* **42**:151 – 162.
doi: http://dx.doi.org/10.1366/0003702884428734.

[8] Grasselli, J. G. (1991) JCAMP-DX, a standard format for exchange of infrared-spectra in computer readable form. *Pure Appl. Chem.* **63**:1781 – 1792.
doi: http://dx.doi.org/10.1351/pac199163121781.

[9] Davies, A. N., and Lampen, P. (1993) JCAMP-DX for NMR. *Appl. Spectrosc.* **47**:1093 – 1099.
doi: http://dx.doi.org/10.1366/0003702934067874.

[10] Lampen, P., Hillig, H., Davies, A. N., and Linscheid, M. (1994) JCAMP-DX for mass-spectrometry. *Appl. Spectrosc.* **48**:1545 – 1552.
doi: http://dx.doi.org/10.1366/0003702944027840.

[11] Lancashire, R. J. (2007) The JSpecView project: an Open Source Java viewer and converter for JCAMP-DX, and XML spectral data files. *Chemistry Central Journal* **1**:31.
doi: http://dx.doi.org/10.1186/1752-153X-1-31.

[12] Cammack, R., Fann, Y., Lancashire, R. J., Maher, J. P., McIntyre, P. S., and Morse, R. (2006) JCAMP-DX for electron magnetic resonance(EMR). *Pure Appl. Chem.* **78**:613 – 631.
doi: http://dx.doi.org/10.1351/pac200678030613.

[13] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Wenger, R. K., Yao, H. Y., and Markley, J. L. (2008) BioMagResBank. *Nucleic Acids Res.* **36**:D402-D408.
doi: http://dx.doi.org/10.1093/nar/gkm957.

[14] Doreleijers, J. F., Nederveen, A. J., Vranken, W., Lin, J. D., Bonvin, A., Kaptein, R., Markley, J. L., and Ulrich, E. L. (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J. Biomol. NMR* **32**:1 – 12.
doi: http://dx.doi.org/10.1007/s10858-005-2195-0.

[15] Wallace, B. A., Whitmore, L., and Janes, R. W. (2006) The Protein Circular Dichroism Data Bank (PCDDB): A bioinformatics and spectroscopic resource. *Proteins-Structure Function and Bioinformatics* **62**:1 – 3.
doi: http://dx.doi.org/10.1002/prot.20676.

[16] Whitmore, L., and Wallace, B. A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers* **89**:392 – 400.
doi: http://dx.doi.org/10.1002/bip.20853.

[17] Murray-Rust, P., and Rzepa, H. S. (2003) Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* **43**:757 – 772.

[18] Murray-Rust, P., Rzepa, H. S., and Wright, M. (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. *New J. Chem.* **25**:618 – 634.
doi: http://dx.doi.org/10.1039/b008780g.

[19] Julian, R. K. (2003) The IUPAC/ASTM Unified Standard for Analytical Data: AnIML. *Scientific Computing* www.scientificcomputing.com/the-iupac-astm-unified-standard.aspx.

[20] Davies, A. N. (2007) Herding AnIMLs (no, it’s not a spelling mistake): Update on the IUPAC and ASTM Collaboration on Analytical Data Standards. *Chemistry International* **29**(6).

[21] Lancashire, R. J., and Davies, A. N. (2006) Spectroscopic Data: The Quest for a Universal Format. *Chemistry International* **28**(1).

[22] Lancashire, R. J. (2000) The use of the Internet for teaching Chemistry. *Anal. Chim. Acta* **420**:239 – 244.
doi: http://dx.doi.org/10.1016/S0003-2670(00)00895-3.