Tools and methods for circular dichroism spectroscopy of proteins: a tutorial review

A. J. Miles, Robert W. Janes and B. A. Wallace

Circular dichroism (CD) spectroscopy is a widely-used method in biochemistry, structural biology and pharmaceutical chemistry. More than 24,000 papers published in the past decade have included CD characterisations of proteins; many of those studies have also included other complementary chemical, biophysical, and computational chemistry methods. This tutorial review describes the background to the technique of CD spectroscopy and good practice methods for high quality data collection. It specifically focuses on both established and new methods and tools available for experimental design and interpretation, data processing, visualisation, analysis, validation, archiving, and accession, including tools developed to enhance the complementarity of this method with other structural and chemical biology studies.

Key learning points
1. Circular dichroism (CD) spectroscopy, a widely-used method for examining the structures and conformational changes of proteins, provides complementary information to that obtainable by other biophysical, chemical, and structural biology techniques.
2. Methods for good practice in measuring, processing, analysing, and interpreting CD spectra of proteins are described.
3. Means of accessing and utilising links to a wide range of online and downloadable tools for comparisons, secondary structure analyses, and predictions of CD spectra are presented.
4. Information is provided on how to access archived CD data sets and associated metadata in the Protein Circular Dichroism Data Bank (PCDDB), and its links to other data bases and validation protocols, and online information.
5. Examples of recent studies and developments utilising CD in novel studies of proteins exemplify its complementarity to other methods.

1. Introduction to circular dichroism spectroscopy of proteins

Biomolecules such as proteins are built up of chiral subunits that produce signals when illuminated by circularly-polarised light in the near and far ultraviolet wavelength ranges where the amide and carbonyl groups of the polypeptide backbones absorb. Circular dichroism (CD) spectroscopy is an optical spectroscopic method which exploits the differential absorption of left- and right-circularly polarised light by such chromophores, and can be harnessed to derive structural information about protein conformations. It has been widely used to discern the secondary structure of proteins based on electronic transitions in the far ultraviolet (UV) wavelength region (≈240 to 170 nm) and to monitor the local tertiary structure environment of aromatic amino acid residues in the near UV region (≈300 to 260 nm), as a function of their physical or chemical environment, amino acid composition (i.e. mutations), or intermolecular interactions.1–4 In the past decade, more than 24,000 papers have been published using CD spectroscopy to characterise the structures of polypeptides and proteins. The information derived by CD spectroscopy, which can also include information on dynamic changes in solution or in environments such as membranes or films, is often complementary to that produced by other biophysical, computational and chemical methods such as crystallography, cryo-electron microscopy, NMR spectroscopy, FTIR spectroscopy, vibrational circular dichroism (VCD) spectroscopy,5 molecular dynamics simulations, and low angle scattering. Indeed, in the same 10 year time period, nearly half of the publications which included CD spectroscopy also included information derived from at least one of these other techniques.

CD spectroscopy has a number of advantages with respect to the higher resolution structural techniques such as crystallography, electron microscopy, and NMR spectroscopy in that it requires relatively small amounts of sample under conditions (temperature, concentration and components present) that may be more comparable to those found in cells. This has resulted in

---

* Institute of Structural and Molecular Biology, Birkbeck University of London, London WC1E 7HX, UK. E-mail: b.wallace@mail.cryst.bbk.ac.uk
+ School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK
its wide-spread use both by the biochemical and structural biology communities to complement the information derived by those other biophysical methods, as well as by the pharmaceutical industry, to assess whether a protein is correctly folded, to monitor structural changes induced by interactions with ligands including other proteins, and to determine protein stability under environmental stresses induced by, for example, changes in pH or temperature. In recent years circular dichroism beamlines have been developed at synchrotron light sources, taking advantage of both their high light flux, which enables faster collection of data from smaller amounts of protein, and the higher information content available due to the lower wavelengths that can be achieved at these high-intensity light sources.

This tutorial review discusses not only methods and software currently available for obtaining, processing, validating, and interpreting high quality CD data obtained using lab-based (commercial) CD instruments (as well as synchrotron radiation circular dichroism (SRCD) beamlines) on a variety of samples types, but also the bioinformatics tools and resources available for determining novel details of the structure and function of proteins based on such data.

2. Methods, tools, and protocols for CD data collection, analyses, and display

2.1 Measuring CD spectra

To optimise the amount of high quality and reproducible CD data obtained from a given sample, it is essential to follow good practice protocols for data collection (see Table 1 for example). To accurately determine the secondary structure of a protein based on CD data, the data obtained must include a spectral range covering, at least, the wavelengths between 240 and 190 nm; more accurate results are obtained if data collected includes even lower wavelengths, because more electronic transitions (peaks) will be included, increasing the information content of the data. To achieve such measurements, conditions must be used so that the total absorbance of the sample (due to protein, buffer and other added components) is below ~1.2 at all wavelengths. This may be challenging for wavelengths below 210 nm, where the absorbance tends to rise, sometimes precipitously, due to the cumulative effects of the buffer and the peptide chromophores along with contributions from water (or other solvents) and the composition material of the optical cell. In addition, light scattering effects from any undissolved protein or suspended particles in the solution, such as lipids present in membrane samples, may also contribute to this challenge.

The overall absorbance of the sample can be monitored by the dynode voltage or high tension (HT) signal produced during data collection. This is a measure of the voltage applied to the detector to amplify the small circular dichroism signal. The maximum HT cutoff values for individual CD instruments differ, but they correspond to the maximum dynode voltage reading...
**Table 1** Workflow for recommended data collection and processing procedures. The column on the right indicates how to avoid potential problems that may arise

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare highly purified protein in low absorbing buffer</td>
<td>≥95% of protein should be the protein of interest</td>
</tr>
<tr>
<td>Based on estimated concentration choose a cell pathlength so that the absorption is &lt;1.2 at all wavelengths</td>
<td>If the absorbance is too high, choose a shorter pathlength cell or a different buffer</td>
</tr>
<tr>
<td>Verify cell pathlength</td>
<td>Cells with pathlengths &lt;0.1 mm can be measured using the interference fringe method</td>
</tr>
<tr>
<td>Measure protein concentration accurately</td>
<td>If possible, measure concentration again immediately before measuring spectrum</td>
</tr>
<tr>
<td>Collect repeat CD spectra of sample and baseline</td>
<td>Monitor the HT signal. Make sure it does not exceed the linear range of the instrument</td>
</tr>
<tr>
<td>Average the sample data</td>
<td>Make sure there are no outliers due to un-equilibrated or leaking or light-sensitive sample or buffer components</td>
</tr>
<tr>
<td>Average the baseline data</td>
<td>Make sure the baseline and sample spectra overlay at wavelengths &gt; 250 nm (where there should not be a protein signal)</td>
</tr>
<tr>
<td>Subtract averaged baseline from averaged sample</td>
<td>Important when comparing spectra measured on different instruments or after the lamp has been changed</td>
</tr>
<tr>
<td>Calibrate the net (sample-baseline) spectrum (optional)</td>
<td>Using the determined values of cell pathlength, protein concentration and mean residue weight</td>
</tr>
<tr>
<td>Scale to standard units</td>
<td></td>
</tr>
</tbody>
</table>

above which the sample absorbance is too high for sufficient light to penetrate. Above this value the apparent CD signal and the intensity of the unpolarised light signal that emerges from the sample decrease to a level where the spectrum becomes noisy, and lead to distortions in both the magnitude and the shape of the measured peaks. Hence the maximum HT value for a given CD instrument needs to be determined.

For a given sample, minimisation of the total absorbance can be achieved by choosing buffer constituents with low absorbance in both the near and far UV wavelength region, and where this is not possible, using the lowest possible concentration of buffer and salts without compromising the stability of the protein. In addition a combination of protein and buffer concentrations, optical cell pathlength, optical cell material, and instrument parameters such as slit width and averaging time can be used to optimise the CD signal.8,10

Secondary structure analyses are also significantly affected by the accuracy of the protein concentration measurement, since this will have an effect on the magnitude of the CD spectrum when scaled to standard units.8,9 The most widely used colorimetric methods for determining protein concentrations, including Biuret, Lowry, bicinchoninic acid assays, and Coomassie blue staining; all produce different values/accuracies for proteins (depending on their amino acid compositions). Measuring the absorption of the sample at 280 nm (λ<sub>280</sub>) is the most convenient and reliable method (although its accuracy can depend on the number of aromatic amino acids and to a lesser extent, their location (surface or buried) in the protein). Such measurements can be achieved (without wasting a great deal material) using micro-UV spectrometers such as Nanodrops.

Another important consideration is if there is a time lag between sample preparation, concentration determination, and CD measurements, or even during the course of a long series of CD measurements, there may be a change in the protein concentration, due to aggregation or precipitation, especially if the sample is unstable or sensitive to light. This issue can be obviated in part by measuring the concentration of the protein immediately before (or as near in time as possible to) measuring the CD spectrum. It can also be monitored by examining the HT measurements obtained during the course of the CD measurements. If the HT values decrease, this could be indicative of protein precipitation/aggregation, bubble formation, or even sample leakage, during the course of the measurement, and should indicate that a new sample needs to be used.

Using an inaccurate value for the optical cell pathlength will also have significant effects on spectral magnitude and therefore the accuracy of secondary structure analysis. This can be an issue when using demountable cells with pathlengths less 0.01 cm, as the manufacturer-reported values can have a wide margin of error. In addition, the loading and assembly of such cells can be non-reproducible.10 However, accurate pathlength measurements can be obtained for these cells using the interference fringe method,10 which requires use of a standard benchtop UV/Vis spectrophotometer.

Finally, as CD instruments are comprised of a number of optical components, there can be variations between instruments which lead to small differences in spectra of the same sample measured on two different instruments. Such disparities can be reduced or mitigated by obtaining calibration measurements with a standard reference material such camphorsulphonic acid (CSA) or ammonium camphour sulfonate (ACS) measured on the same instrument used for the sample measurements.7,11,12 Both of these compounds have two well-defined and well-separated peaks of known absolute magnitude, so comparisons between
the calculated and the experimental values can be used to create a calibration curve over the wavelength range of the spectrum. Multiplying the measured CD values of the protein spectrum by the CD values of this curve at each wavelength, will adjust the spectrum so that it better matches the spectra measured on any other instrument where this procedure is carried out.

2.2 Processing methods and tools

CD spectra are usually produced by averaging a set of repeat scans (or accumulations) obtained for the sample, after which the averaged spectrum of the baseline (a solution that contains all components of the sample except the protein), is subtracted. The final net spectrum can then be calibrated against a solution (CSA or ACS) with a known signal, and normalised to standard units of mean residue ellipticity (MRE, degrees cm$^2$ dmol$^{-1}$ residue$^{-1}$) or delta epsilon ($\Delta \varepsilon$, M$^{-1}$ cm$^{-1}$) using the concentration, optical cell pathlength, and mean residue weight value for the protein. Each step in this process provides an opportunity to identify anomalies in the data. For example, there may be a systematic change in the data as a function of time that is apparent in repeated scans (possibly as a result of light-induced conformational changes, or because the protein precipitates/aggregates with time). Processing data using a standard spreadsheet is possible but this method can be time-consuming; hence there are a number of possible alternatives described in this section.

2.2.1 Instrument-associated software. CD instruments are supplied with their own data processing software packages, which enable basic processing procedures such as averaging of scans (accumulations), baseline subtraction, scaling to standard units and data smoothing. These functions are found in the “spectra manager” for Jasco instruments and the “trace manipulation menu” in the Pro Data viewer provided with the Chirascan spectrophotometer. For both the Aviv and Olis instruments, they are in the data processing menu. All manufacturers also provide a comprehensive set of mathematical functions, ranging from simple arithmetic such as scaling and normalisation at a peak wavelength to curve fitting. The Olis software also includes software for basic dataset analyses using Singular Value Decomposition (SVD) methodology, also available in CDtool software. This is a form of principal component analysis for a set of spectra, for example from a temperature or time series, in which the spectral features are reduced to a small number of principal components or basis spectra that, in different proportions, can be summed to reproduce each experimental spectrum in the series. The coefficients that determine the contribution of a particular basis spectrum can be plotted as a curve that represents the change in secondary structure during the course of the experiment. SVD analyses provide alternatives to monitoring the CD signal at a single peak wavelength to identify changes, and have the advantage of incorporating information from the entire wavelength range of the spectrum rather from just a single wavelength (see Fig. 1).

Instrument-processed data can be exported in ASCII format for external use including secondary structure analysis using the tools described in Section 3. However there is no consensus of file structure which would enable comparisons of data produced on different instruments, for example, data collected on a benchtop instrument and at an SRCD beamline, or two different bench-top instruments, which would enable simple comparisons. For this reason, the generic data processing software CDToolX, which is described in the following section, was created for use with the output of any CD or synchrotron radiation circular dichroism (SRCD) instrument, and provides formatted output results that are instrument-independent.

2.2.2 CDToolX. CDToolX, and its predecessor CDTool, are freely-available, downloadable tools designed for data processing and for some types of analyses of protein CD spectra. CDToolX is designed for use on Windows 10 operating systems (or with Windows 10 emulators on other operating systems), whereas CDTool works on earlier operating systems such as Windows 7 and Windows XP (which may often be installed on older CD instruments) or Mac-based Windows emulators. At present, CDToolX is compatible with files created by the following instruments (and possibly others): Aviv, Jasco, (.txt file outputs only), Applied Photophysics, Chirascan (.cdnn format), and outputs by most SRCD beamlines. In addition, an ASCII text file with two tab-delimited columns containing wavelength and CD data (i.e. those generated by spreadsheets), and generic files (.gen) produced by the original CDTool program can be used. This software was specifically designed to enable identification of data anomalies before they
are incorporated into the final processed spectrum, a strategy that provides a complete data trail for archiving and regulatory purposes. Additionally, cross instrument comparison can be enhanced by calibration performed using spectra of CSA or ACS, which adjusts for differences in instrument polarisation. The simple math operations supported include smoothing, spectral magnitude scaling and scaling of multiple spectra to each other at selected wavelengths as a sensitive means of detecting spectral differences in cases where the concentrations of the protein in the samples has not/cannot be accurately measured (Fig. 2). SVD of datasets can also be performed (Fig. 1) and the data can be stored in a local MySQL database to be retrieved in the CDToolX plot window in units of Δε or mdeg, either calibrated or uncalibrated.

2.2.3 Summary. Instrument manufacturers provide a comprehensive set of instrument-specific tools for processing, analysis and data manipulation, whereas CDTool14 and CDToolX13 are dedicated to the basic tasks required by a majority of users. The latter also provide a transparent user-friendly environment, which enables comparisons of data between instruments, and facilitates CSA calibration to adjust for differences in instrument polarisation. Furthermore CDToolX13 is the primary processing software used at a number of SRCD beamlines around the world.

2.3 Methods and tools for analysing protein secondary structure from CD data

Protein CD spectra include two types of electronic transitions arising from peptide absorptions in the far UV. These include an n → π* transition that gives rise to a signal at ~222 nm, and parallel and perpendicular π → π* transitions that give rise to signals at lower wavelengths in the range from ~190–210 nm (illustrated in Fig. 3). In addition, an intra-amide charge transfer transition arising from through-space interactions occurs at around 180 nm, which is generally only accessible using synchrotron radiation as a light source. The different types of secondary structures that are adopted by polypeptide chains (for example, helix or beta strand) depend on the dihedral (Φ, Ψ) angles between adjacent residues in the polypeptide chain; the magnitudes and wavelengths of the CD signals that arise from the electronic transitions depend upon these angles and therefore upon the secondary structure of the protein. The overall shape and magnitude of a protein CD spectrum in the far ultraviolet (UV) wavelength region reflects a linear combination of its constituent secondary structural elements. Fig. 4 shows examples of spectra of proteins that are each dominated by one type of secondary structure, but most proteins contain multiple types of secondary structures that all contribute to their net CD spectrum.

There are a number of deconvolution methods available for obtaining quantitative secondary structural information from CD spectra,1,2,14 each of which requires a reference dataset of CD spectra produced from proteins with known (crystal) structures, in order to produce a calculated secondary structure spectrum that best matches the query (experimental) spectrum. These methods range from relatively simple least squares
algorithms, to more complex ridge regression (RR) and singular value deconvolution (SVD) methods with variable selection (VS) functions, which fit the experimental spectrum to a weighted sum of the individual reference spectra. In general, a broader-base (more, and more varied) of components present in the reference database, allows more accurate identification of the protein structure that produced the query spectrum. However, the results of some types of analyses can be optimised by the use of specialised datasets, such as those designed for integral membrane proteins (which, due to their presence in low

dielectric environments, tend to have transitions at somewhat different wavelengths than soluble proteins\(^\text{2,13}\)). The characteristics and component types of current publically-available datasets are listed in Table 2, along with an indication as to which type of protein each database is best suited for.

Alternative analysis methods\(^\text{15}\) use neural networks (NN) trained on sets of CD reference data. The accuracy of these depend upon both the suitability of the dataset used for training, on the breadth of protein spectra available, and the spectral wavelength range covered.

### 2.3.1 The DichroWeb analysis server\(^\text{16}\)

DichroWeb is a freely-available online resource for determining protein secondary structures based on CD and SRCD spectra. It provides five different calculation algorithms [CONTINLL\(^\text{17}\) (a regression restraint method), SELCON3,\(^\text{18}\) CDSSTR\(^\text{19}\) and VARSLEC,\(^\text{20}\) (all SVD methods with variable selection), and K2D\(^\text{15}\) (a neural network method now superseded by the stand-alone K2D3 method – see Section 2.3.3)]. These methods can be used with the reference datasets noted above, except for K2D, which uses its own reference data, and VARSLEC which has a built-in (unchangeable) dataset.

The dataset chosen for use will depend on the type of protein to be analysed. For example, to analyse the spectrum of a highly structured soluble protein, an appropriate choice may be SP175,\(^\text{21}\) which contains 71 high quality spectra of a bioinformatics-defined set of soluble globular proteins, covering all protein fold classes, which has a selectable wavelength range covering 175 nm to 240 nm. Alternatively, the SMP180\(^\text{22}\) dataset contains not only spectra of the soluble globular SP175\(^\text{21}\) proteins, but also 29 membrane protein spectra with low wavelength cutoffs of 180 nm, and hence is more suitable for analysing membrane proteins. The choice of dataset may be restricted by the low wavelength cutoff of the experimental data: for data which extend to a low wavelength of only 190 nm, truncated versions of SP175\(^\text{21}\) and SMP180\(^\text{22}\) are available. Spectral data that do not extend to wavelengths at least as low as 190 nm or below are not suitable for analysis by DichroWeb,\(^\text{16}\) as they do not have sufficient information content to enable detailed definitions of secondary structure.\(^\text{21}\) The secondary structure definitions used to create these reference datasets are those defined by the “Dictionary of Protein Secondary Structures” (DSSP) algorithm,\(^\text{23}\) based on conformational characteristics identified in crystal structures.

The DichroWeb\(^\text{16}\) server accepts file formats produced by most commercial and SRCD instruments, plus CDTool\(^\text{X13}\) and CDTool\(^\text{14}\)-generated files, and simple two column (wavelength, value) text files. User files can be directly uploaded to the online server, and a number of parameters can be manually selected, including the high and low wavelengths of the data, the lowest wavelength to be considered in the analysis (which may differ from the lowest wavelength collected, if that wavelength resulted in an HT value that exceeded the cutoff limit of the instrument), and the spectral units, either mdeg or mean residue ellipticity (MRE). Input of the wavelength step size is also required. The algorithm and dataset to be used are selected from a dropdown box, and if the input spectrum has units of mdeg, the concentration in mg ml\(^{-1}\), optical pathlength in cm, and the protein mean residue weight are also required inputs.
so that the spectrum can be converted to MRE units prior to analysis. The output pages (Fig. 5a and b) provide both a compact and an extended listing of the secondary structural results, and a plot of the back-calculated spectrum based on the secondary structure determined, which is overlaid on the experimental spectrum, as a visual indicator of the quality of the result (i.e. correspondence between the shapes and magnitudes of the calculated and measured spectra). The compact results table simply lists the predicted secondary structure fractions, whereas the extended results table (Fig. 5a) also provides a goodness-of-fit parameter known as the normalised root mean squared deviation.

### Table 2

Characteristics of currently-available reference datasets for use in CD analyses, and the servers on which they are available. The final column indicates the types of proteins for which the dataset may be best suited. The individual protein spectral data comprising the SP175\(^{21}\) and SMP180\(^{22}\) datasets can be downloaded from the PCDDB\(^{35}\).

<table>
<thead>
<tr>
<th>Reference dataset</th>
<th>Wavelength range (nm)</th>
<th>Number of proteins</th>
<th>Server</th>
<th>Types of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET1(^{25})</td>
<td>178–260</td>
<td>29</td>
<td>DichroWeb(^{36})</td>
<td>soluble, globular</td>
</tr>
<tr>
<td>SET2(^{25})</td>
<td>178–260</td>
<td>22</td>
<td>DichroWeb</td>
<td>soluble, globular</td>
</tr>
<tr>
<td>SET3(^{25})</td>
<td>185–240</td>
<td>37</td>
<td>DichroWeb</td>
<td>soluble, globular</td>
</tr>
<tr>
<td>SET4(^{25})</td>
<td>190–240</td>
<td>43</td>
<td>DichroWeb</td>
<td>soluble, globular</td>
</tr>
<tr>
<td>SET5(^{25})</td>
<td>178–260</td>
<td>17</td>
<td>DichroWeb</td>
<td>soluble, globular</td>
</tr>
<tr>
<td>SET6(^{25})</td>
<td>185–240</td>
<td>42</td>
<td>DichroWeb</td>
<td>soluble, globular and denatured proteins</td>
</tr>
<tr>
<td>SET7(^{25})</td>
<td>190–240</td>
<td>48</td>
<td>DichroWeb</td>
<td>soluble, globular and denatured proteins</td>
</tr>
<tr>
<td>SMP175(^{21})</td>
<td>175–240</td>
<td>71</td>
<td>DichroWeb</td>
<td>soluble, globular (bioinformatics definitions) and membrane and soluble proteins</td>
</tr>
<tr>
<td>SMP180(^{22})</td>
<td>180–260</td>
<td>129</td>
<td>DichroWeb</td>
<td>soluble, globular especially β-sheet</td>
</tr>
<tr>
<td>SP175+29</td>
<td>175–240</td>
<td>79</td>
<td>BeStSel(^{29})</td>
<td>soluble, globular especially β-sheet</td>
</tr>
</tbody>
</table>

Fig. 5 (a) Example results page obtained using the DichroWeb\(^{36}\) server for a “good quality analysis”. (top) The protein name (ATPase, PCDDB\(^{35}\) code (CD0004003000)) is displayed and the analysis method used [ContinLL15] is listed on the next line. Below this is the NRMSD\(^{24}\) “goodness-of-fit parameter”, which should optimally be \(<0.1\) (as it is in this example), indicating a close correspondence between the back-calculated and measured spectra. If it is not, then another method, reference data set and/or scale factor should be used. (middle) Tables [shaded areas] of calculated secondary structure results obtained using the CONTINLL\(^{27}\) method and the appropriate (SMP180,\(^{22}\) membrane protein) reference data set. The arrow at the right of the top row indicates what is usually the closest/most suitable solution. The lower shaded box indicates other possible solutions obtained using other types of calculations. (bottom) Plot showing a comparison of the experimental spectrum (crosses), the back-calculated closest match spectrum (stars), and the difference spectrum (vertical bars) between the experimental and back-calculated spectra (vertical lines). The low NRMSD\(^{27}\) is consistent with the close match of the calculated and experimental spectra. These, plus the small magnitude difference spectrum indicate this is a “good quality” analysis. (b) Example results page obtained using the DichroWeb\(^{36}\) server for a “poor” quality analysis. This was obtained for an intrinsically disordered protein, HASPA (PCDDB code: CD0005282000). As in Fig. 5a, except in this case neither the (high) NRMSD\(^{27}\) value (>0.1) nor the correspondence between the calculated and experimental spectra, suggest that the best solution is an accurate reflection of the secondary structure. This is because this is an intrinsically disordered protein and is comprised of mostly unordered or disordered (not helical, sheet nor turn) secondary structures. As such it does not have well-defined phi, psi angles, and as the reference dataset does not contain many spectra of proteins with significant amounts of disorder (largely because this type of protein does not tend to crystallise), the NRMSD\(^{27}\) value is high. There is also a greater difference between the experimental and back-calculated spectra for this protein as compared to that for the well-ordered protein depicted in the Fig. 5a.

---

\(^{24}\) Normalised root mean squared deviation.

\(^{25}\) viewpoints.

\(^{26}\) DichroWeb.

\(^{27}\) Normalised root mean squared deviation.

\(^{28}\) Free online tools for the determination of protein secondary structure from circular dichroism data.

\(^{29}\) BeStSel.
Tutorial Review

which is an indication of the correspondence between the measured data and the back-calculated spectrum produced from the derived secondary structures, and is similar to an ‘R-factor’ in crystallography. It is defined as:

\[
NRMSD = \sqrt{\frac{\sum_i (\theta_{\\text{exp}} - \theta_{\\text{calc}})^2}{\sum_i (\theta_{\\text{calc}})^2}}
\]

where \(\theta_{\\text{exp}}\) and \(\theta_{\\text{calc}}\) are the experimental and back-calculated ellipticities, respectively, at each wavelength in the spectrum. Analyses that produce high NRMSD values (\(>0.1\)) or poor correspondences between the experimental and calculated best fit spectra are strong indications that the protein of interest includes features not present in any of the proteins present in the reference data base.

The calculated results (Fig. 5) include the following secondary structure types: regular and distorted alpha helix, regular and distorted beta sheet, and turns.\(^{25}\) The distorted helix and sheet fractions include the residues at either end of an alpha helix and one residue at either end of a beta strand, which have slightly different dihedral angles than the corresponding canonical structures based on crystallographic data and thus have slightly different characteristic spectra. The ‘turn’ fraction includes beta turns, bends, and bridges as defined by DSSP.\(^{23}\) All other types of structure, including random coil, are classified as ‘unordered’. Two exceptions to these classifications are those produced by structure, including random coil, are classified as ‘unordered’. Two exceptions to these classifications are those produced by

...
Table 3 (a) Examples of secondary structure percentages calculated using different methods of analyses and different reference datasets for soluble proteins, compared with the values obtained from their crystal structures. For this comparison, all spectra were analysed in the wavelength range from 240 nm to 190 nm, and only the percentages of total helix and total sheet are indicated. The last two rows (for hemerythrin) demonstrate the effect of spectral magnitude (scaling factor of 1.0 (penultimate line) vs. 1.3 (final line)) on the NRMSD24 goodness-of-fit parameter (column 1), and the calculated secondary structures (columns 3–6). For the correct scaling (1.3 in this case), the resulting NRMSD24 is smaller than that for the incorrectly scaled data, and the calculated structure more closely matches secondary structure of the protein, as calculated from the crystal structure (column 7) using DSSP23
(b) Example of secondary structure analyses for a mostly disordered protein (the hydrophilic acylated surface protein (HASPA) from L. major) using different calculation methods and reference data bases. As is typical for intrinsically disordered proteins, there is no crystal structure available for it, so the values given in the last column are average values calculated from several bioinformatics prediction sites (SPOT-1D, RaptorX, NetSurfP). 'Other' indicates disordered secondary structure (i.e. neither helix nor sheet). Its spectrum (Fig. 3) clearly indicates (from its shape) that the protein is mostly disordered; however since there are no fully-disordered proteins in the PDB30 and hence in the CD reference datasets, even the "best" results produced a relatively poor (high) NRMSD24 values, and only modest matches to the bioinformatics predictions

2.4 Identifying spectral artifacts and improving CD spectra

Quality control considerations for both the integrity of proteins under study and the analysis methods used to characterise them are important for producing suitable CD analyses. This has led to the development of ValiDichro34 software that can be used to identify some artifacts and irreproducibility in CD data, and/or errors in the parameters used for calculating secondary structures from CD data.

2.4.1 The ValiDichro server34. The ValiDichro server34 provides a means of testing CD data and associated metadata as a guide to good practice and to aid publication. As well as the dedicated standalone server, there is a version embedded in the PCDDB, which is used to vet depositions before they are included as entries in the database. Data quality tests are based on the common characteristics of CD protein spectra observed in the literature and are given a designation of a pass (P), a flag (F) for small deviations from the norm, or a fail (X) if the deviation is more severe.

ValiDichro34 tests CD spectral data quality, reproducibility, and completeness (wavelength range), as well as the associated metadata for consistency. Quality tests include maximum and minimum peak magnitudes (outside normal ranges for spectra when scaled to units of Δε or mean residue ellipticity). Outliers may indicate concentration or optical cell pathlength errors, or problems with the sample such as the flattening of peaks due to
Instrument-specific formats (saved as ASCII files), and two- and three-column free formats, where the third column containing HT data is available. Some metadata are automatically extracted from the files and the rest are manually entered via text boxes and dropdown lists. The output includes a pass/flag/fail status for each test and suggestions designed to improve the data.

2.5 Public archive of CD spectra (the PCDDB\textsuperscript{15})

In 2006, with CD and (SRCD) having become ubiquitous methods employed in the molecular life sciences, there was an obvious need for a public repository where authors could store data and make it available to other researchers. Following extensive public consultations with the structural biology, spectroscopic and bioinformatics communities, the full resource, described below, was created and has evolced since then to include additional links to other databases, bioinformatics resources, and parameters for quality control assessments. To date, more than one million files have been downloaded, either as individual files or the full database contents.

2.5.1 The PCDDB\textsuperscript{15}. The Protein Circular Dichroism Data Bank (PCDDB) is a freely-accessible web-based repository of CD and SRCD data inspired by the availability of other online resources such the UNIPROT\textsuperscript{36} and PDB\textsuperscript{30} databanks. Publishers are increasingly requiring that CD and SRCD data included in an article be made publically available to ensure good practice and traceability of the data, and these are functions which the PCDDB\textsuperscript{15} fulfils. The PCDDB\textsuperscript{15} site includes many links to other tools and tutorial videos (see Section 4) demonstrating protocols and procedures for the collection and analysis of CD data.

Depositions to the databank require depositor registration although accessions and downloads do not. When a deposition is to be made, the depositor requests an appropriate number of PCDDB ids (nine digit accession codes) be reserved. The spectra can be uploaded as ASCII text files, as generated by most CD instruments, or as CDToolX.gen or CDTool.gen files. Although only the fully-processed spectrum is required, spectra from all stages of the data processing, including the CSA or ACS spectrum used for instrument calibration, can be uploaded for completeness/good practice/tracing. Information including experimental conditions (including protein purity), instrument parameters, and protein sequence data are essential; links to other sites, including UNIPROT\textsuperscript{36} and the PDB,\textsuperscript{30} and other parameters such as enzyme classification, and citation details for the article in which data is presented, are strongly encouraged in order to create a comprehensive entry. Spectra are tested by a version of Vali Dichro\textsuperscript{34} (see Section 2.4.1) and given a pass/flag/fail status before they can be committed to the database. Depositors can download a summary ValiDichro\textsuperscript{34} report that can be submitted to journals along with their manuscripts as an indication of data quality and accessibility.

The PCDDB\textsuperscript{15} provides an extensive list of searching criteria, including all of the above fields. The full contents of the database or just a single spectrum can be downloaded as .gen files, or as a two column (wavelength, CD) .pcd file.

2.6 Methods for comparing CD spectra

Spectra can be visually compared using CDToolX\textsuperscript{13} and CDTool\textsuperscript{14} software (see Section 2.2), in generic spreadsheet...
2.7 Predictions of CD spectra

A number of methods have been developed for the prediction of a CD spectrum based on a protein’s atomic coordinates. This can be useful for comparing two proteins when, for example, the X-ray diffraction, cryo-electron microscopy, or NMR structure of one is available, but only the CD spectrum of the other is available. Such comparisons may be useful to confirm homology, to determine if a mutant has folded correctly (and in a similar way to the wild-type protein), and for monitoring the effects of ligand binding or different environmental factors on conformation.

The ab initio method of DichroCalc uses quantum mechanical calculations based on the averaged crystal structure to generate predicted spectra, whereas the PDB2CD and PDBMD2CD servers generate CD spectra using proteins present in a reference set with similar structural characteristics, by least squares fitting procedures. Both the PDBMD2CD server (based on structures produced using user-input molecular dynamic simulation results), and the downloadable programme SESCA generate predicted spectra using a principal component approach.

2.7.1 The DichroCalc server. DichroCalc employs a matrix method, which simplifies the otherwise computationally extensive and challenging quantum mechanical calculations required to determine how a polypeptide absorbs circularly polarised light, based on its structural coordinates. PDB files can be uploaded to the server either as a single file or multiple files archived in .zip, .Z, .gz, or .tgz formats. Alternatively the PDB code of the protein can be entered, with multiple codes separated by commas. The user is offered a choice of which chromophores to use in the calculation; by default these are the backbone chromophores only, but the inclusion of the backbone charge transfer transitions can be beneficial for far UV calculations, albeit at the expense of significantly increased computation time. Analyses of the near UV wavelength range require the addition of aromatic side chain transitions of phenylalanine, tyrosine and tryptophan. Calculations that include the side chain chromophores of asparagine and glutamine residues, as well as peptide bonds between lysine and aspartic acid that create cyclic peptides, are also available. The backbone chromophores are essential for calculating far UV spectra but any combination of the other three selections can be added to the mix. Each chain of a multi-subunit protein is analysed separately. The results are delivered as two column (wavelength, CD) text files.

2.7.2 The PDB2CD server. The algorithm underlying the PDB2CD server develops the CD spectrum of the query protein from the CD spectra of structurally-similar proteins in either the SP17521 or SMP18022 datasets. It achieves this using three different levels of structure-based information in the associated PDB files. The first level is the secondary structure content based on the DSSP definitions of alpha helix (H), beta-strand (E) and the other (O) category (which includes everything else). Then the localized topological features of the secondary structure components are considered by assessing the relative juxtaposition of secondary structural elements. Finally the overall structural similarity between the query and derived subset is determined. Following further refinement, the CD data of the remaining similar proteins are averaged to obtain the CD spectrum of the query protein.

The user is required to choose between reference datasets SP17521 or SMP18022 before uploading a PDB file or entering...
4. identifying different structural populations in the input data if

the PDB\(^{30}\) code. In the output, the calculated spectrum is
accompanied by a summary of the calculated secondary structure
and a link to its PCDB\(^{36}\) entry, if available. The website displays
the spectrum calculated by PDB2CD\(^{38}\) superimposed on the
experimental spectrum (Fig. 7).

2.7.3 The PDBMD2CD server\(^{39}\). Following the development
of the PDB2CD\(^{38}\) server, it became apparent that there was
a demand for a similar tool for predicting CD spectra from multiple
input coordinate files from a variety of sources such as molecular
dynamics (MD) simulation structures, NMR ensemble structures,
and multiple homologous proteins. MD is a method of exploring
conformational space of proteins using the laws of motion to
provide dynamic and thermodynamic information. The results are
difficult to experimentally validate, however one such method is
to compare the predicted CD spectra of multiple structures to
experimental spectra.

Given coordinate data as input, the PDBMD2CD\(^{39}\) server
creates basis spectra representing seven secondary structural
types derived from a least squares regression of 83 spectra
present in the SMP180 reference data set. A second set of basis
spectra is also calculated from the PDB\(^{39}\) structures of proteins
in the reference set with the closest secondary structure content
to the query protein. A predicted spectrum is then derived
from each basis set and the two are averaged to produce the
calculated spectrum.

PDB\(^{30}\) format files can be uploaded to the server as archive
files (.tar/.gz/.zip/.bz2) or as 4 digit PDB\(^{30}\) codes separated by
commas. The input page also has an optional function ‘Split
NMR models’, which can be selected when analysing multiple
NMR structures, so that each structure can be considered
individually. When the calculations are complete, the output
is divided into three tabs: results, clustering and compare to
experiment. The first of these displays a plot of all the predicted
spectra and the averaged predicted spectrum with an interactive
3D representation of the most representative input structure.
Other information produced includes the average RMSD
between all the generated spectra and the average spectrum.
The clustering tab displays k-mean clustering, a method of
identifying different structural populations in the input data if
>50 structures have been predicted. Finally, for the ‘compare to
experiment tab’, an experimental spectrum can be uploaded as

a two column text file so that it can be compared with the
predicted spectra (Fig. 7).

2.7.4 SESCA software\(^{40}\). The structure-based empirical
calculation approach, SESCA, works in a somewhat similar manner
to the PDBMD2CD\(^{39}\) software. In SESCA, a choice of reference datasets
is available, with each using different structural assignments to
calculate a set of between three and eight basis spectra. The
predicted spectrum is then obtained using a weighted average of
the basis spectra derived from secondary structural information
obtained from the PDB\(^{30}\) file of the target protein. This software
requires downloading and installation in a Python environment.

2.7.5 Comparison of prediction algorithms. Whilst spectral
predictions by DichroCalc agree in a general way with experi-
mental spectra, especially for the spectral magnitudes of the 190
and 208 nm peaks, it is less accurate in assigning a value at
222 nm, leading to distortions in the overall shape of the spectra
of helical proteins. This is not an issue for the spectra of beta-sheet
rich proteins, which generally do not have a peak at 222 nm.
PDB2CD\(^{38}\) produces more complete spectral matches to experi-
mental data, but it requires similar proteins to the query protein
be present in the comparison dataset in order to work accurately.
PDBMD2CD\(^{39}\) and SESCA\(^{40}\) both enable prediction of CD spectra
from multiple coordinate files as well as single structure files.

3. Applications demonstrating uses of
CD spectroscopy for studying proteins

The above-described wide array of tools and resources now
available for different types of analyses are now enabling novel
applications of CD spectroscopy to answer specific biological ques-
tions that can complement other structural biology techniques.\(^{5}\)
Some recent illustrative studies include the following selected
examples which focus on the cross-overs/complementarity with
other methods, as noted in Section 1.

In an investigation of environmental effects on structure
and stability,\(^{41}\) the secondary structures and thermal stabilities
of the voltage-gated sodium channel NavMs (sodium channel
from Magnetooccus marinus) in different environments were
undertaken using CD and thermal melt CD studies. CD data was
processed with CDTool\(^{14}\) and analysed using the DichroWeb\(^{16}\)
server. SVD analysis of thermal melt data was performed using
the CDToolX.\(^{13}\) Different amphipols and detergents were examined
in order to identify amphipathic environments that stabilised
the protein structure; indicating these could be suitable for cryo-
electron microscopy studies.

In a study aimed at identifying drug binding sites\(^{42}\) not
visible by crystallography, also for the NavMs channel, thermal
melt circular dichroism spectroscopy was used to compare the
stability of the full-length channel and a pore-only construct in
the presence and absence of the anticonvulsant drug valproic
acid. Deconvolution analyses and secondary structure determina-
tions for apo- and drug-bound forms of both types of constructs at
different temperatures using the DichroWeb\(^{16}\) server indicated
that the interaction involved the pore-only domain, and not the
voltage-sensor region, an unexpected and novel result for this type
of drug. This study provided structural data in a system for which high resolution methods such as crystallography and cryo-electron microscopy have not yet been able to identify the binding sites.

The pharmaceutical industry relies on CD to ensure consistency between batches of drugs both during development and production. Perez et al.,\textsuperscript{43} characterised a novel monoclonal antibody by near UV CD, far UV CD, and fluorescence spectroscopies. Secondary structure analyses using the DichroWeb\textsuperscript{16} server, after minimizing the differences in protein concentration between batches by scaling, then enabled comparisons using DichroMatch.\textsuperscript{37}

Another example is a study by Zheng et al.\textsuperscript{44} which illustrated the utility of predictive (computational) tools for evaluating the conformational preferences of the C-terminal peptide of the P66 domain of human immunodeficiency virus HIV-1 reverse transcriptase. CD spectroscopy was then used to confirm that the structure did indeed form a beta-sheet rich structure, as predicted, and the resulting spectrum compared favourably with the theoretical CD spectra generated by the computational server PDB2CD.\textsuperscript{38}

4. Websites and downloadable tools for processing, analysing, interpreting, and comparing CD spectroscopic and other structural biology data

Data repository
The Protein Circular Dichroism Data Bank (PCDDB),\textsuperscript{35} a databank of validated CD spectra and metadata, is accessible at: https://pcddb.cryst.bbk.ac.uk. It is an archive of published spectroscopic data, data collection parameters, and auxiliary information for each protein, and includes the amino acid sequence (including any construct-specific differences), the Enzyme Classification code (E. C. number) – if relevant, the CATH structure classification (if available), and cross-references to the corresponding entries in the Uniprot and Protein Data Banks.

Tools for processing and validation of CD data
CDToolX\textsuperscript{13} and its predecessor CDTool\textsuperscript{14} are generic (instrument-independent) programmes for data processing and singular value deconvolution analyses which are downloadable at http://www.cdtools.cryst.bbk.ac.uk.

A video describing how to process data using CDToolX\textsuperscript{14} can be found on YouTube at: https://www.youtube.com/watch?v=ajNkf9OzBU

Videos describing how to set up and use the CDToolX\textsuperscript{14} database are available on YouTube at:
https://www.youtube.com/watch?v=8Doy77Uuo-s
https://www.youtube.com/watch?v=u8Afbbmt6Jg
A detailed video describing how the SVD function in CDToolX\textsuperscript{13} can be used is available on the YouTube: https://www.youtube.com/watch?v=lymfg4hiyf4

The ValiDichro server\textsuperscript{34} which enables checking of spectra and metadata for quality and validity at https://pcddb.cryst.bbk.ac.uk/validichro/

Tools for secondary structure analyses
The DichroWeb server\textsuperscript{16} for secondary structure analyses is available at http://DichroWeb.cryst.bbk.ac.uk/html/home.shtml and includes a wide range of computational options, including reference databases, methods, and analysis tools.

The BeStSel server,\textsuperscript{29} is accessible at http://bestsel.elte.hu/index.php for secondary structure analyses, and is especially focused on beta sheet-rich proteins.

The K2D3 secondary structure server,\textsuperscript{32} based on a neural network approach, is accessible at http://cbdm-01.zdv.uni-mainz.de/~andrade/k2d3/

The CONTINLL\textsuperscript{17} secondary structure analysis program is downloadable at http://s-provencher.com/contin.shtml

The SESCA analysis program\textsuperscript{40} requires a Python environment, and is downloadable at https://www.mpibpc.mpg.de/cesca

Reference data sets for secondary structure analyses
The component spectra of the SP175 dataset\textsuperscript{21} and associated secondary structure and other metadata are downloadable from the PCDDB\textsuperscript{35} (see above) as accession codes: CD0000001000 to CD0000071000.

The component spectra of the SMP180 dataset\textsuperscript{22} (for analysis of membrane proteins) are downloadable from the PCDDB.\textsuperscript{35} This data set includes 29 membrane proteins (accession codes CD0000099000 to CD0000128000), the SP175\textsuperscript{21} proteins (see above), and 26 additional soluble proteins (accession codes CD0000072000 to CD0000098000).

Examples of tools for spectral prediction and comparisons
The DichroMatch server\textsuperscript{37} for identifying similar protein structures on the basis of related spectra is available at: https://pcddb.cryst.bbk.ac.uk/dichromatch.php

The PDB2CD server\textsuperscript{38} for calculation of theoretical CD spectra for proteins based on their PDB coordinates is accessible at: https://pdb2cd.cryst.bbk.ac.uk

The PDBMD2CD server\textsuperscript{39} for calculation of spectra based structures derived from molecular dynamics simulations is accessible at: https://pdbmd2cd.cryst.bbk.ac.uk

The DichroCalc server\textsuperscript{13} for calculation of theoretical spectra from protein structures, is accessible at: https://comp.chem.nottingham.ac.uk/dichrocalc/index.html

Instrument-specific software tools
CD instruments are supplied with software for processing and analysing data collected on their specific instrument. Software manuals and/or tutorials can be accessed at:

Jasco training videos are available at: https://jascoinc.com/training-video/video-category/spectra-manager-for-cd/

The APP Chirascan manual can be requested at: https://www.photophysics.com

The Olis Instruments website homepage is at: http://olisweb.com/

YouTube videos about CD procedures and software
Informational and training videos are available on the “PCDDB Channel” at: https://www.youtube.com/user/ThePcddb/videos?app=desktop&view=0&sort=dd&shelf_id=1
5. Conclusions

This tutorial review has described (and illustrated) methodologies for the collection and analysis of CD data. It includes extensive information on tools and resources available for analyses, validation, comparison and interpretation of CD spectroscopy data and the interoperability of CD spectroscopy with other structural biology techniques.

Conflicts of interest

There are not conflicts of interest to declare.

Acknowledgements

This work was supported by grants from the UK Biotechnology and Biological Sciences Research Council (BBSRC) Bioinformatics and Biological Resources Fund (most recently, BB/P024092 to BAW and BB/P024106 to RWJ).

References

5 L. A. Nafie, Chirality, 2020, 32, 667.
41 S. M. Ireland, A. Sula and B. A. Wallace, Biopolymers, 2018, 109, e23067.