Summary of Validation Methods

The ultimate aims of high quality data collection and spectral processing procedures are to produce reproducible spectra which reflect as many of the structural features of the protein as possible, and remove as many of the causes of variation in spectral shape and magnitude as can be achieved. The result should be that the same polypeptide under the same conditions measured at different times and places on different instrumentations will present identical spectra (1,2). A number of the quality tests have been based on common characteristics observed for protein CD spectra in the literature. Variations from standard characteristics are not always indicative of problems, but are worthy of further investigation or consideration. Indeed, these may be due to particularly interesting and novel spectral features. Some tests will produce both Flag (F) and Fail (X) results, and some only one of these two outcomes. The end of each test description notes which type of indicators are produced by that test.

Tests for Data Completeness

Missing Wavelengths: To assure there are no missing data points in the spectrum, the differences between sequential wavelengths are assessed relative to the most common wavelength interval found in the spectrum. Faults of this nature often result from human error when trimming or otherwise processing the data, or transferring between software or spreadsheets. (X)

Wavelength Range: The qualities of secondary structure analyses derived from CD data (3) are dependent upon the amount of data available; most analysis programs require data at least between 190 and 240 nm. If more data are available in the low wavelength area, this can improve the quality of the analyses (4), and the availability of data up to 280 nm generally improves the definition of the baseline alignment. The minimal standard for a Pass result for this test is a wavelength range between 205 nm and 255 nm, a region containing a significant portion of the spectral features generated by the peptide bond. If only a narrower range is obtained, the user is advised to find a different set of experimental conditions, such as changing pathlength, concentration, and/or buffer conditions. (F)

Wavelength Interval: The standard wavelength interval used in most CD experiments and analysis software is 1 nm. If the interval is larger, this can distort the shape of spectral features. Intervals shorter than 1 nm are acceptable. This procedure also checks that the wavelength interval stated in the metadata is that same as that present in all of the spectral data (except for the calibration spectrum). i.e. The Sample Spectrum interval should match the Baseline Spectrum interval. (X)

Tests for Metadata and Spectral Data Consistency

Metadata:

UniProt Sequence: This tests if the amino acid sequence provided by the user (minus any expression tags stated) matches with sequence(s) associated with the UniProt code(s) provided. (F)

Number of Residues: The number of residues listed in the metadata is compared with amino acid sequence provided. (X)

Mean Residue Weight: The mean residue weight provided (necessary for unit conversion) is compared with that calculated from the molecular weight and the number of residues provided by the user. (X)
**Molecular Weight:** The molecular weight (in Daltons) is calculated from the amino acid sequence provided in the metadata and compared with the value of the molecular weight provided. (X)

**Spectral Data:**

**Final Processed Spectrum:** The Final Processed Spectrum is calculated from the Net Spectrum, the Calibration Spectrum and the metadata provided (if these are provided) and compared with the Final Processed Spectrum provided, including the magnitude (unit conversion). (X)

**Average Sample or Baseline Spectrum:** The Average Spectrum from multiple individual Raw Sample and/or Baseline Spectra provided is calculated and compared with the Average Spectra provided. (F)

**Excess Smoothing:** The Average or Raw Spectrum is compared with the Net Smoothed Spectrum. If a spectrum has been over-smoothed (too large a smoothing interval) (5), the magnitudes of the peaks will appear truncated relative to the unsmoothed spectrum or the positions of the peaks will be shifted. If the peaks in the Net Smoothed Spectrum differ from those in the Average or Raw Spectrum by >5% then a fail is generated. (X)

**Experimental Temperature:** This test is to check for potential decimal point errors, or possible use of the wrong units. The acceptable range is -10 to 99°C. (X)

**Tests for Quality**

**Minimum Peak Size:** If the maximum size (absolute value) of the highest peak, in units of Delta Epsilon, is not at least 1.0, this is indicative of a calculation error or that the wrong sample and baseline files have been used. (X)

**Maximum Magnitude:** The magnitude at each wavelength throughout the spectrum (over the wavelength range from 178 to 245 nm) in units of Delta Epsilon is compared against an envelope of maximum and minimum values found within a curated reference set of over 150 validated CD spectra deposited in the PCDDB) which cover secondary structure and fold space (6). This procedure may either detect errors which occurred during unit conversion, or may be indicative of an interesting feature. In latter case a Flag does not necessarily mean an error, but may be a novel characteristic worthy of the author’s attention. (F)

**Noise (Spectral features at 260-270 nm):** This procedure assesses the magnitude (in units of Delta Epsilon) of data points between 260 nm and 270 nm in the Final Processed Spectrum. If two or more successive values exceed +/-0.25 Delta Epsilon, this may be worthy of investigations because in general, protein circular dichroism do not generate signals larger than this at these wavelengths. Deviations from zero can possibly be attributed to too noisy data, or errors in baseline matching with sample spectra. (F)

**Calibration (CSA/ACS Peak Ratio):** The peak ratio listed in the metadata for the calibration standard (either CSA or ACS) is compared with the ratio present in the Calibration Spectrum, if provided. In addition, if the ratio varies from the standard ratio of 2.0 by more than 10%, the value is flagged. (F)

**Maximum HT Voltage:** The High Tension (HT) or High Voltage (HV) is a measure of the degree of amplification by the photomultiplier tube, and the maximal values determined empirically for different types of instruments (depending on their definition and way of measuring this). Exceeding of the maximal value will tend to depress the magnitude of the measurement. (F,X)
Concentration/Pathlength Relationship: There is an inverse relationship between the likely optimal ratio between pathlength and concentration. As a rough guide, the following equation can be used to estimate the optimum concentration (mg/ml) from the pathlength (cm) provided: \( y = \frac{x}{0.01} \), where \( y \) = concentration (in mg/ml) and \( x \) = pathlength (in cm). Few examples of values outside this range have been found by assessment of the curated spectra present in the PCDDB. (F)

Flat Topped Peaks: Too high absorbances in a sample can result in CD peaks above a certain magnitude not being properly measured, resulting in flat peaks (in the wavelength dimension) above that value. CD and HT spectra are assessed to determine if >4 successive points in a peak have the same value. (X)

Feature Width: Narrowing of peaks, especially at low wavelengths, has been associated with poor signal-to-noise ratios. If the width of peaks at half maximum is less than 10 nm, a flagged result is generated. This criterion was decided upon through consultation with experts in the field and assessed using curated spectra present in the PCDDB. (F)

Peak Locations: The wavelength locations of positive and negative peaks present in curated spectra present in the PCDDB are compared with the peak locations present in the Final Processed Spectrum. If peaks are discovered at unexpected locations, a flagged result is generated and the location of the peak is listed. This may be an indication of an interesting spectral feature rather than a quality issue. (F)

Standard Deviation at Peak: The wavelength locations of peaks are identified for each of the raw repeats to examine instrument and sample stability. The standard deviation between the locations of these peaks is calculated, and if this value exceeds 1.5 nm a flagged result is generated. (F)

HT Voltage in the 240 to 260 nm Region: HT Spectra in the wavelength region from 260 nm to 240 nm, are typically flat or have a negative slope, unless there is a large absorption due to the buffer components. The HT values are normalised between 0 and 1 and the average gradient between these wavelengths is then assessed. An increase of more than 0.05 will generate a flagged result. (F)

Standard Deviation: To check reproducibility of the measurements, if the standard deviation of measurements between repeats of the Raw Spectrum exceeds 5% of the of the value found at three or more consecutive wavelengths and exceeds 0.6 millidegrees then a flagged result is generated and the wavelengths of concern are noted. This method has been tested using curated data in the PCDDB (6). (F)

Projection: This test looks to see if the spectra bear resemblance to known CD spectral characteristics. Basis spectra (the 5 eigenvectors with the highest eigenvalues) are generated from reference sets of high quality reference database spectra (7,8). These basis spectra can be thought of as the 5 shapes that when added together in various ratios can recreate the shape in the Final Processed Spectra. The degree to which the shape of the submitted spectrum differs from the closest shape of the spectrum created from the eigenvectors is flagged if the value exceeds a flagging limit. Due to variations in shape between the spectra of membrane and soluble proteins different reference sets (7) and flagging limits are used for each. This test may indicate novel spectral features rather than a quality issue. (F)


