eSPC, an Online Data Analysis Platform for Molecular Biophysics

ChiraKit 1.0
User Documentation

October 2024

This document is a work in progress. Some information may be incomplete or subject to change.

Table of Contents

1. Import data

1.1. Input file

ChiraKit can parse many different types of files:

A) A comma-separated-values **(.csv)** file with a header, and two or three columns. In the case of two columns, the first contains the wavelength and the second the CD signal. In the case of three columns, the first contains the wavelength, the second the sample name, and the third the CD signal. In both cases the first column must be called 'wavelength'.

	Standard	Standard	Standard
1	wavelength	time	CD_signal
2	220	1	10
3	220	2	20
4	220	3	30
5	220	4	40

Figure 1. Example of a valid CSV format.

B) A csv file **(.csv)** with a header and **more than three columns**: 'wavelength', CD signal of the first sample , ..., CD signal of the n-th sample.The first column must be called 'wavelength' or 'xydata'.

	Standard	Standard	Standard	Standard
1	wavelength	CD_signal_A	CD_signal_B	CD_signal_C
2	220	1	2	3
3	221	1	2	3
4	222	1	2	3
5	223	1	2	3

Figure 2. Example of a valid CSV format. Each column represents one CD curve.

C) A csv file **(.csv)** with header lines containing the metadata. The metadata keys and values should be separated by ':'. The CD data should start two lines after a line containing the column heading of 'Wavelength' and should be arranged in columns. The first column has the wavelength data (in nm), and the subsequent columns have the spectral data (e.g., scan repeats). This type of file is exported by the ProData software from AppliedPhotophysics™.

Available Dimensions:	2		
Wavelength	Wavelength: 180nm - 290nm	Step Size: 1nm	E
Repeat	2 repeats in set.	-iter option selected	
Available Properties:	1		
CircularDichroism			
Data:			
CircularDichroism			
Wavelength	Repeat		
	0	1	
290	-0.313281	-0.284654	
289	-0.353384	-0.46977	
288	-0.488774	-0.488609	
287	-0.277936	-0.486426	

Figure 3. Example of a ProData generated CSV.

D) A PCDDB file (**.pcd**, Protein Circular Dichroism Data Bank format) with header lines containing the metadata. The metadata keys and values are organised into two columns, with the second column starting after the 60th character. The CD data starts after the first line containing 'DATA' and 'wavelength', and is arranged in columns. The first column has the wavelength data (in nm), and the subsequent columns have the spectra data (e.g., 1. Wavelength. 2. Final. 3. HT. 4.Smoothed. 5. Avg. Sample. 6. Avg. Baseline.).

```
Publication Journal                                         Bioinformatics
Publication Title                                           A reference database for circu
Publication Volume                                          22
Publication Pages                                           1955-1962
Depositor Name                                              Andrew Miles
Department/School name                                      Crystallography
University/Institution/Corporation                          Birkbeck
Depositor Country                                           United Kingdom
Name of Principal Investigator (if not depositor)           B.A. Wallace
DATA (1. Wavelength. 2. Final. 3. HT. 4.Smoothed. 5. Avg. Sample. 6. Avg. Baseline.)
280.0     8.98482E-02       3.84900E+02       8.75000E-02       1.63333E+00       1.50000E+00
279.0    -1.28364E-02       3.84000E+02      -1.25000E-02       1.60000E+00       1.56667E+00
```

Figure 4. Example of a PCDDB file. Original source:
https://pcddb.cryst.bbk.ac.uk/deposit/CD0000003000&dl=1#

E) A 'generic' file (**.gen**, CDTool format) with header lines containing the metadata. The metadata keys and values are organised into two columns, separated by tabs. The CD data starts after the first line with more than three tab divided columns. The first column has the wavelength data (in nm), and the subsequent columns have the spectral data (e.g., 'Final Processed Spectrum', 'HT values (from raw spectrum 1)'). Duplicated or missing data is removed.

```
Smoothing window        5
PDB                     1ed9
Swissprot               P00634
Code                    Lees, J.G., Wien, F., Miles, A.J. & Wallace, B.A., 2006 Bioinformatics A refere
structure space 22 1955-1962
279.0       5.62969E-02     1.53648E+00     5.62969E-02     0.00000E+00     0.00000E+00     0.00000E+00
278.0       2.21733E-01     1.53303E+00     2.21733E-01     0.00000E+00     0.00000E+00     0.00000E+00
277.0       4.74813E-02     1.52912E+00     4.74813E-02     0.00000E+00     0.00000E+00     0.00000E+00
```

Figure 5. Example of a 'generic' file. Original source: https://pcddb.cryst.bbk.ac.uk/deposit/CD0000002000&dl=1#

F) A text file (**.dat**) with header lines containing the metadata. The metadata keys and values should be separated by ':'. The CD data starts two lines after the line containing 'MDCDATA' in it and is arranged in columns. The first column has the wavelength data (in nm), and the subsequent columns have the spectral data. There could be one or more scans separated in blocks. The voltage data is read from the column called 'CD_Dynode'.

```
Averaging Time   : 1.000 seconds
Settling Time    : 0.333 seconds
Multi-Scan Wait  :  1.00 seconds

$DATA
$MDCNAME:Scan_#1
$MDCDATA:1:14:2:3:4:9
 X  CD_Signal  CD_Error  CD_Current_(Abs)  CD_Delta_Absorbance  CD_Dynode  Jacket_Temp.
280.000  0.000  0.047  1.013  0.000  242.6  19.99
279.000  -0.006  0.083  1.013  -0.000  242.9  19.98
278.000  0.044  0.148  1.013  0.000  243.2  19.99
```

Figure 6. Example of a valid text file. Original source: https://github.com/samirelanduk/CDtool/blob/master/ftests/files/three-aviv-baseline.dat

G) Aarhus synchrotron data file (**.dx, where x means any number**) with header lines containing the metadata. The metadata keys and values are separated into two columns where the second column starts after the 32th character (or the character ':'). The CD data starts after the first line containing 'Lambda' and 'CD', and is arranged in columns. The first column has the wavelength data (in nm), the second column has the spectral data ('CD/mdeg'), the fifth column the voltage data ('Servo_Volts'), and the eighth column the temperature data ('temperature').

```
;Grating / CD slit              LEG / 08.75
;  Comments:
;  Sample0: Water
;  Cell type: AS 121a123
;  330-170nm 1nm 20av 1sc Slit 8.75mm
;Lambda         CD/mdeg         Y_Comp./mdeg    DC_Bias         Servo_Volts     Z_Motor         Beam_current    temperature
330.000         1.4818          0.0095          6.82440         4.37090         0.3505          180.3981        25.00
329.000         1.0437          0.0095          6.82440         4.36770         0.2660          180.3248        25.00
```

Figure 7. Example of a CD data file generated by the AU-SRCD facility at ASTRID2.

H) A csv file (**.csv**) containing a line with the words 'ORIGIN' and 'JASCO'. The CD data is arranged in columns and the column names are defined between the lines 'XUNITS' and 'FIRSTX'. This type of file is exported by Jasco™ circular dichroism instruments.

XUNITS	NANOMETERS		
YUNITS	CD [mdeg]		
Y2UNITS	HT [V]		
Y3UNITS	ABSORBANCE		
FIRSTX	260		
LASTX	180		
NPOINTS	801		
FIRSTY	-0,28448		
MAXY	0,86205		
MINY	-8,74658		
XYDATA			
260	-0,284475	254,707	0,758667
259,9	-0,279734	254,753	0,759227

Figure 8. Example of a CD data file exported from a Jasco™ circular dichroism instrument.

I) A csv file **(.csv)** with header lines containing the metadata. The metadata keys and values should be separated by ':'. The CD data starts two lines after the line containing 'Wavelength' and 'Temperature' in it and is ordered in columns. The first column has the wavelength data (in nm), and the subsequent columns have the spectral data (one column per temperature). The associated temperature data is available one line before the CD data starts. This type of file is exported by the ProData software from AppliedPhotophysics™.

1.2. Parameters for molar ellipticity / extinction

Each input file can be associated with a certain set of experimental parameters (Box '2. Parameters for molar ellipticity / extinction') . The expected units to input in ChiraKit are:

Molecular weight:	Dalton (g/mol)
Number of chromophores:	Unitless
Concentration:	mg/ml
Path length:	millimetre

These parameters are useful to compare different samples. Indeed, the CD signal can be normalised based on the sample concentration (C), cell path length (L), and molecular weight (M) or mean unit molecular weight (M_u).

A summary of the CD units is presented below.

Name	Formula	Units
Differential absorbance	$A_l - A_r$	Unitless

Differential milliabsorbance	Absorbance / 1000	Unitless
Molar extinction	Absorbance * M / (C*L)	Lts * mol^{-1} * cm^{-1}
Mean unit molar extinction (Δε)[1]	Absorbance * M_u / (C*L)	Lts * mol^{-1} * cm^{-1} * $chromophores^{-1}$
Degrees	Absorbance * 32.98	Degrees (º)
Millidegrees	Degrees / 1000	mº
Molar ellipticity	Degrees * M * 100 / (C*L)	º*cm^2 * $dmol^{-1}$
Mean unit molar ellipticity[1]	Degrees * M_u * 100 / (C*L)	º*cm^2 * $dmol^{-1}$ * $chromophores^{-1}$

Table 1. Units of CD Measurement. A_l and A_r refer respectively to the left- and right-circularly polarised light. 'chromophores' refers to the total number of chromophoric units. For measurements of polypeptides in the far-UV region (185 - 240 nm), the number of peptide bonds is used for normalisation, because the amide bond is the main contributor to the CD signal.[1]

1.3. Processing

Given a selection of spectra (column 'Spectrum/a 1' inside the box '3. Processing'), we can:

1) Subtract another spectrum (column 'Spectrum 2')
2) Add another spectrum (column 'Spectrum 2')
3) Smooth them using a Savitzky-Golay filter (window size of 3, 6, 8 or 10 nm)
4) Average them
5) Average them in batch mode
6) Zero them by subtracting the mean signal within a selected wavelength range interval (3, 5, 10, 20 or 40 nm)

The 'subtract', 'sum', 'smooth' and 'zero' operations will produce as many spectra as defined in the column 'Spectrum/a 1'. The 'average' operation will generate one spectrum. The 'batch average' operation depends on the selected 'N'. For instance, if we have six spectra and select 'N' equal to 3, we will generate two new spectra.

Regarding the 'subtract' and 'sum' operations, the data must have been measured with the same wavelength step and wavelength range. If not, a linear interpolation can be activated. In that case, 1) the resulting spectrum will cover the region where the spectra overlap and 2) a linear interpolation will be performed, implying that the

[1] Kahn, Peter C. "[16] The interpretation of near-ultraviolet circular dichroism." Methods in enzymology. Vol. 61. Academic Press, 1979. 339-378.

new spectrum will be computed at all the wavelength data points present in both spectra.

The following behaviour is applied for the high tension voltage curves. For the subtract, addition, smooth or zero operations, the generated spectra will have the same high tension voltage curve as the spectra selected in the column 'Spectrum/a 1'. For the (batch) average operation, the high tension voltage curves will be averaged.

2. Analysis

2.1. Creating a dataset

When generating a dataset for the thermal /chemical unfolding analysis, custom analysis, or spectra comparison module, the CD signal will be merged based on the wavelength data. In other words, wavelengths present in all spectra will be used. Non-shared wavelengths will be discarded.

2.2. Thermal unfolding

2.2.1. Equilibrium two-state model (from monomer to tetramer)[2,3]

We assume that the protein only exists in the native (folded) or unfolded state and that there is an equilibrium between these two states given by the unfolding reaction $F_n \leftrightarrows nU$. The signal is described by the following equation:

$$Signal(T) = f_N(k_N T + b_N) + f_U(k_U T + b_U)n \quad (1)$$

where f_N and f_U are the folded and unfolded fractions, T is the temperature in Kelvin units, b_N and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the temperature dependence.

The concentration of [U] and [F] depend on the equilibrium unfolding constant K_u:

$$K_u(T) = \frac{[U]^n}{[F_n]} \qquad (2)$$

[2] Santoro, M. M., & Bolen, D. W. (1988). Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl. alpha.-chymotrypsin using different denaturants. *Biochemistry*, *27*(21), 8063-8068.

[3] Bedouelle, H. (2016). Principles and equations for measuring and interpreting protein stability: From monomer to tetramer. *Biochimie*, *121*, 29-37.

where $K_u(T)$ is a function of the thermodynamic parameters T_m (temperature at which $K_u(T)$ equals one), ΔH (enthalpy of unfolding), and ΔCp (heat capacity of unfolding):

$$K_u(T) = e^{-\Delta G \,/\, R_{gas} T} \quad (3)$$

$$\Delta G = \Delta H(1 - \frac{1}{T_m}) - \Delta Cp(T_m - T + Tlog(\frac{T}{T_m})) \qquad (4)$$

where R_{gas} is the gas constant. For monomers, f_N and f_U can be calculated from $K_u(T)$ only, while for oligomers, the protein concentration is also required.

2.2.2. Equilibrium three-state model (with a monomeric intermediate)

We assume that the protein exists in the native (folded), intermediate or unfolded state and that there is an equilibrium between these three states given by the reactions $F_n \leftrightarrows nI$ and $I \leftrightarrows U$.[4] The signal is described by the equation:

$$Signal(T) = f_N(k_N T + b_N) + f_U(k_U T + b_U)n + f_I(b_I)n \qquad (5)$$

where f_N, f_I and f_U are the fraction of protein in the folded intermediate and unfolded states, T is the temperature in Kelvin units, b_N, b_I and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the temperature dependence (of the folded and unfolded states).

The concentration of [U], [F] and [I] depend on the unfolding constant $K_1(T)$ and $K_2(T)$:

$$K_1(T) = \frac{[I]^n}{[F_n]} \qquad (6)$$

$$K_2(T) = \frac{[U]}{[I]} \qquad (7)$$

where $K_x(T)$ (x = 1 or 2) is a function of the thermodynamic parameters T_x (temperature at which $K_x(T)$ equals one), ΔH_x (enthalpy of unfolding), and $\Delta C_{p,x}$ (heat capacity of unfolding):

[4] Bedouelle, Hugues. "Principles and equations for measuring and interpreting protein stability: From monomer to tetramer." *Biochimie* 121 (2016): 29-37.

$$K_x(T) = e^{-\Delta G_x / R_{gas} T} \qquad (8)$$

$$\Delta G_x = \Delta H_x (1 - \frac{1}{T_x}) - \Delta C_{p,x}(T_x - T + T log(\frac{T}{T_x})) \qquad (9)$$

where R_{gas} is the gas constant. For monomers, f_N, f_I and f_U can be calculated from $K_1(T)$ and $K_2(T)$ only, while for oligomers, the protein concentration is also required.

For monomers, $\Delta C_{p,x}$ are assumed to be zero, or $\Delta C_{p,1}$ is fitted globally based on the total $\Delta C_{p,th}$ given by the user ($\Delta C_{p,th} = \Delta C_{p,1} + \Delta C_{p,2}$). For dimers, $\Delta C_{p,x}$ are assumed to be zero, or $\Delta C_{p,1}$ is fitted globally based on the total $\Delta C_{p,th}$ given by the user ($\Delta C_{p,th} = \Delta C_{p,1} + 2\Delta C_{p,2}$). For trimers and tetramers, $\Delta C_{p,th}$ are assumed to be zero.

2.2.3. Equilibrium three-state model (with a n-meric intermediate)

We assume that the protein exists in the native (folded), intermediate or unfolded state and that there is an equilibrium between these three states given by the reactions $F_n \leftrightharpoons I_n$ and $I_n \leftrightharpoons nU$.[5] The signal is described by the equation:

$$Signal(T) = f_N(k_N T + b_N) + f_U(k_U T + b_U)n + f_I(b_I) \qquad (10)$$

where f_N, f_I and f_U are the fraction of protein in the folded intermediate and unfolded states, T is the temperature in Kelvin units, b_N, b_I and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the temperature dependence (of the folded and unfolded states).

The concentration of [U], [F] and [I] depend on the unfolding constant $K_1(T)$ and $K_2(T)$:

$$K_1(T) = \frac{[I_n]}{[F_n]} \qquad (11)$$

$$K_2(T) = \frac{[U]^n}{[I_n]} \qquad (12)$$

[5] Bedouelle, Hugues. "Principles and equations for measuring and interpreting protein stability: From monomer to tetramer." *Biochimie* 121 (2016): 29-37.

where $K_x(T)$ (x = 1 or 2) is a function of the thermodynamic parameters T_x (temperature at which $K_x(T)$ equals one), ΔH_x (enthalpy of unfolding), and $\Delta C_{p,x}$ (heat capacity of unfolding):

$$K_x(T) = e^{-\Delta G_x / R_{gas} T} \qquad (13)$$

$$\Delta G_x = \Delta H_x (1 - \frac{1}{T_x}) - \Delta C_{p,x}(T_x - T + T log(\frac{T}{T_x})) \qquad (14)$$

where R_{gas} is the gas constant. For monomers, f_N, f_I and f_U can be calculated from $K_1(T)$ and $K_2(T)$ only, while for oligomers, the protein concentration is also required.

For monomers, $\Delta C_{p,x}$ are assumed to be zero, or $\Delta C_{p,1}$ is fitted globally based on the total $\Delta C_{p,th}$ given by the user ($\Delta C_{p,th} = \Delta C_{p,1} + \Delta C_{p,2}$). For dimers, $\Delta C_{p,x}$ are assumed to be zero, or $\Delta C_{p,1}$ is fitted globally based on the total $\Delta C_{p,th}$ given by the user ($\Delta C_{p,th} = \Delta C_{p,1} + \Delta C_{p,2}$). For trimers, $\Delta C_{p,th}$ are assumed to be zero.

2.2.4. Irreversible two-state model

We assume that the protein only exists in the native (folded) or unfolded state and that the unfolding reaction is irreversible (N → U).[6] The signal is described by the equation:

$$Signal(T) = f_u(k_U T + b_U) + f_n(k_N T + b_N) \qquad (15)$$

where T is the temperature, k_N, b_N are the slope and intercept of the pre-transition baseline (native), and k_U and b_U are the slope and intercept of the post-transition (unfolded) baseline.

The fraction of natively folded molecules f_n as a function of temperature (T) is given by the differential equation:

$$\frac{\partial f_n}{dT} = \frac{-k(T) f_n}{v} \qquad (16)$$

where v is the scan rate (in degrees per minute), and $k(T)$ is the rate constant (min^{-1}). The rate constant can be derived from the modified Arrhenius formula:

[6] Mazurenko, Stanislav, et al. "Exploration of protein unfolding by modelling calorimetry data from reheating." *Scientific reports* 7.1 (2017): 16321.

$$k(T) = e^{\left(\frac{-Ea}{R}\left(\frac{1}{T}-\frac{1}{T_f}\right)\right)} \qquad (17)$$

where R is the gas constant, Ea is the activation energy, and T_f is the temperature where $k(t)$ equals one. f_n is assumed to be one at the lowest measured temperature.

2.2.5. Irreversible three-state model

We assume that the protein exists in the native (folded), intermediate or unfolded state. The first step is reversible and the second one is irreversible (N ⇆ I → U). The signal is described by the equation:

$$Signal(T) = f_u(k_U T + b_U) + f_n(k_N T + b_N) + f_i(b_I) \qquad (18)$$

where T is the temperature, k_N, b_N are the slope and intercept of the pre-transition baseline (native), k_U and b_U are the slope and intercept of the post-transition (unfolded) baseline, and b_I is the baseline of the intermediate state. f_u, f_n and f_i are respectively the unfolded, folded and intermediate fractions, which are calculated as:

$$\frac{\partial f_u}{dT} = \frac{-k(T)K(T)(1-f_u)}{v(K(T)+1)} \qquad (19)$$

$$f_n = \frac{(1-f_u)}{1+K(T)} \qquad (20)$$

$$f_i = 1 - f_n - f_u \qquad (21)$$

where v is the scan rate (in degrees per minute), and $k(T)$ is the rate constant (min^{-1}). The rate constant can be derived from the modified Arrhenius formula (equation 12). $K(T)$ is the equilibrium unfolding constant (N ⇆ I) and can be calculated as in equation 3 (ΔCp is assumed to be zero).

2.2.6. Local and global parameters

All curves are simultaneously fitted, constrained to share the same values for the thermodynamic parameters (e.g., $\Delta H, T_1$, Ea, T_f). The baseline and slopes are allowed to vary, and users also have the option to set the slopes (k_u , k_n) to zero.

2.2.7. Fitting errors

The standard deviation of all fitted parameters is computed using the square root of diagonal values from the fit parameter covariance matrix reported by scipy.curve_fit function. These values are an approximation (**underestimation**) of the real errors. Relative errors are calculated as $100 * (std(\widehat{\theta}) / \widehat{\theta})$ where $\widehat{\theta}$ refers to the estimate of the parameter and $std(\widehat{\theta})$ to the standard deviation.

2.3. Chemical unfolding (and the linear extrapolation model)

2.3.1. Equilibrium two-state (from monomer to tetramer)

The equilibrium two-state model assumes a linear dependence of stability on the denaturant concentration.[7,8] The signal is expressed by the following equation:

$$Signal(D) = f_N(D)(k_N D + b_N) + f_U(D)(k_U D + b_U)n \quad (22)$$

where f_N and f_U are the folded and unfolded fractions, D is the denaturant concentration in molar units, b_N and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the linear dependence of the signal on D.

The concentration of [U] and [F] depend on the equilibrium unfolding constant K_U :

$$K_u(D) = \frac{[U]^n}{[F_n]} \quad (23)$$

$$K_u(T) = e^{-\Delta G(D) / R_{gas} T} \quad (24)$$

where T is the temperature in Kelvin units and R_{gas} is the gas constant. Assuming a linear dependence of stability on the denaturant concentration, $\Delta G(D)$ can be expressed as

$$\Delta G(D) = M(D50 - D) \quad (25)$$

where $D50$ is the denaturant concentration at which 50% of the molecules are folded and M the slope $\frac{\partial \Delta G}{\partial D}$.

[7] Pace,C.N. and Hermans,J. (1975) The stability of globular protein. *CRC Crit. Rev. Biochem.*, **3**, 1–43.
[8] Myers,J.K., Pace,C.N. and Scholtz,J.M. (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.*, **4**, 2138–2148.

For monomers, f_N and f_U can be calculated from $K_u(D)$ only, while for oligomers, the protein concentration is also required.

2.3.2. Equilibrium three-state (with a monomeric intermediate)

We assume that the protein exists in the native (folded), intermediate or unfolded state and that there is an equilibrium between these three states given by the reactions $F_n \leftrightharpoons nI$ and $I \leftrightharpoons U$.[9] The signal is described by the equation:

$$Signal(D) = f_N(k_N D + b_N) + f_U(k_U D + b_U)n + f_I(b_I)n \qquad (26)$$

where f_N, f_I and f_U are the fraction of protein in the folded intermediate and unfolded states, D is the denaturant agent concentration in molar units, b_N, b_I and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the linear dependence of the signal on D (of the folded and unfolded states).

The concentration of [U], [F] and [I] depend on the unfolding constant $K_1(D)$ and $K_2(D)$:

$$K_1(D) = \frac{[I]^n}{[F_n]} \qquad (27)$$

$$K_2(D) = \frac{[U]}{[I]} \qquad (28)$$

where $K_x(D)$ (x = 1 or 2) is a function of the "thermodynamic" parameters $D50_x$ (concentration at which $K_x(D)$ equals one) and M_x (equal to $\frac{\partial \Delta G_x}{\partial D}$):

$$K_x(D) = e^{-\Delta G(D)_x / R_{gas} T} \qquad (29)$$

where T is the temperature in Kelvin units and R_{gas} the gas constant.

For monomers, f_N, f_I and f_U can be calculated from $K_1(D)$ and $K_2(D)$ only, while for oligomers, the protein concentration is also required.

2.3.3. Equilibrium three-state (with a n-meric intermediate)

[9] Bedouelle, Hugues. "Principles and equations for measuring and interpreting protein stability: From monomer to tetramer." *Biochimie* 121 (2016): 29-37.

We assume that the protein exists in the native (folded), intermediate or unfolded state and that there is an equilibrium between these three states given by the reactions $F_n \leftrightarrows I_n$ and $I_n \leftrightarrows nU$.[10] The signal is described by the equation:

$$Signal(D) = f_N(k_N D + b_N) + f_U(k_U D + b_U)n + f_I(b_I) \qquad (30)$$

where f_N, f_I and f_U are the fraction of protein in the folded intermediate and unfolded states, D is the denaturant agent concentration in molar units, b_N, b_I and b_U are the baseline terms, and k_N and k_U are the slopes to take into account the linear dependence of the signal on D (of the folded and unfolded states).

The concentration of [U], [F] and [I] depend on the unfolding constant $K_1(D)$ and $K_2(D)$:

$$K_1(D) = \frac{[I_n]}{[F_n]} \qquad (31)$$

$$K_2(D) = \frac{[U]^n}{[I]} \qquad (32)$$

where $K_x(D)$ (x = 1 or 2) is a function of the "thermodynamic" parameters $D50_x$ (concentration at which $K_x(D)$ equals one) and M_x (equal to $\frac{\partial \Delta G_x}{\partial D}$):

$$K_x(D) = e^{-\Delta G(D)_x / R_{gas}T} \qquad (33)$$

where T is the temperature in Kelvin units and R_{gas} the gas constant.

For monomers, f_N, f_I and f_U can be calculated from $K_1(D)$ and $K_2(D)$ only, while for oligomers, the protein concentration is also required.

2.3.4. Local and global parameters

All curves are simultaneously fitted, constrained to share the same values for $D50$ (or $D50_{1/2}$) and M (or $M_{1/2}$). The baseline and slopes are allowed to vary, and users also have the option to set the slopes (k_u , k_n) to zero.

2.3.5. Fitting errors

[10] Bedouelle, Hugues. "Principles and equations for measuring and interpreting protein stability: From monomer to tetramer." *Biochimie* 121 (2016): 29-37.

See Section ‘2.2.7. Fitting errors’.

2.4. Custom analysis

2.4.1 The fitting function

The CD signal can also be analysed as a function of any experimental parameter. In this case, the fitting function is given by the user. The following rules are required:

1) The given experimental parameter should be included in the fitting function string. For example, if the experimental parameter name is ‘T’, then using ‘T*z’ as a fitting function implies that we fit the value of ‘z’ from the curve of the CD signal against the values of ‘T’.

2) The function parameters can not contain mathematical characters; e.g., instead of 'dG_h2o', use 'dG_water'.

3) To use exponential, logarithm or root-square functions, you need to write ‘e^(...)’, ‘log(...)’, or ‘sqrt(...)’. The parenthesis are obligatory.

4) Parameters with the pattern ‘Global’, e.g., ‘TmGlobal’ will be shared across all curves.

5) Function parameters with the pattern ‘Pos’ are constrained to be greater than zero. The opposite holds for the parameters with the pattern ‘Neg’.

2.4.2. Initial parameter estimates

The initial estimates for the function parameters are obtained through a log-spaced grid search. The number of combinations to be tested is fixed, but the limits of the search space can be given by the user. The default behaviour is to explore values between 10^{-3} and 10^{3}. The combination of initial values resulting in the lowest residual sum of squares is selected.

2.4.3. Local and global parameters

When fitting the CD signal against the experimental parameter values, all curves will be simultaneously fitted. The constraint is that they share the same values for parameters with the pattern ‘Global’. As for all other parameters, they can vary independently for each curve.

2.5. Curve fitting and fitting errors

The curve fitting procedure is done by using the non-linear least squares method (scipy.curve_fit[11]). The standard deviation is computed using the square root of diagonal values from the fit parameter covariance matrix. For each parameter, the relative errors (in percentage) are computed as the quotient between the parameter standard deviation and the parameter value multiplied by one hundred.

2.6. Comparison of spectra

To comprehend if there are differences between groups of spectra, ChiraKit allows computing averages and the associated standard deviations ('Module 2e. Spectra Comparison', in the app). Moreover, every possible difference spectrum and its associated standard deviation is also computed (through error propagation). ChiraKit provides the euclidean distances between all spectra and a box plot figure with the intra and inter group distributions.

To compare the shapes of spectra and discard the influence of differences in absorbance intensity, the L2 normalisation can be used. This results in the calculation of normalised euclidean distances as defined in the Equation 2 from Oyama, Taiji, *et al* (2022)[12].

3. Spectra decomposition

To take advantage of all the spectral information, instead of monitoring a single wavelength, the loaded spectra can be decomposed into a set of basis spectra.

The CD values of a given spectrum are decomposed as a linear combination of basis spectra as follows:

$$CD(x) = c_1(x)\phi_1 + c_2(x)\phi_2 + c_3(x)\phi_3 + \ ...\ + \ c_n(x)\phi_n \qquad (34)$$

where each ϕ_i is a basis spectrum and c_i the associated coefficients. The basis spectra are orthogonal to each other and have unit norm. The variable x represents the dimension used for the measurement, such as temperature.

To find the set of basis spectra, we can apply singular value decomposition (SVD) directly on the CD data matrix, or SVD on the centred CD data matrix (equivalent to Principal component analysis)[13].

---

[11] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html
[12] Oyama, Taiji, *et al.* "Performance Comparison of Spectral Distance Calculation Methods." Applied Spectroscopy 76.12 (2022): 1482-1493.
[13] Kim, Donggun, and Kisung You. "PCA, SVD, and Centering of Data." arXiv preprint arXiv:2307.15213 (2023).

3.1. Singular value decomposition (SVD)

SVD consist of factorizing the data matrix M of size *m x n* into the matrices U, Σ and V^T such that

$$M = U \Sigma V^T \qquad (35)$$

where U is an *m x m* orthogonal matrix, Σ is an *m x n* non-negative rectangular diagonal matrix, and V^T is an *n x n* orthogonal matrix. U contains the left singular vectors (basis spectra), V^T the right singular vectors and Σ the singular values.

The amount of explained variance associated to each singular value (s_i) is given by the following equation:

$$explained\ variance(s_i) = s_i^2 \ / \ \sum_{j=1}^{n} s_j^2 \qquad (36)$$

The associated coefficients for the i-th basis spectrum and the j-th acquired spectrum is defined as:

$$\phi_{svd,i} \bullet M_j \qquad (37)$$

where M_j is the j-th column (spectrum) of the matrix M.

3.2. Principal component analysis (PCA)

PCA is performed through eigendecomposition of the covariance matrix C:

$$C = VLV^t \qquad (38)$$

where V is the matrix containing the eigenvectors (basis spectra) and L is the diagonal matrix with the eigenvalues (sorted in descending order). The amount of explained variance associated to each eigenvalues (λ_i) is given by the following equation:

$$explained\ variance(\lambda_I) = \lambda_i \ / \ \sum_{j=1}^{n} \lambda_j \qquad (39)$$

It is important to remember that here we refer to the variance of the centred matrix. The associated coefficients for the i-th basis spectrum and the j-th acquired spectrum is defined as:

$$\phi_{pca,i} \bullet M_{meaned,j} \qquad (40)$$

where $M_{meaned,j}$ is the j-th column (spectrum) of the centred matrix M.

3.3. Basis spectrum inversion

Any given basis spectrum ϕ_i can be inverted. As a result the explained variance will remain constant, the new spectrum ϕ_i' will be equal to $(-\phi_i)$ and the weighting function c_i' will be equal to $(-c_i)$.

3.4. Change of basis

Given a set of two or three basis spectra, we can rotate them to obtain a new set of basis spectra with the following properties: the first new basis spectrum is similar to the first acquired spectrum (e.g., lowest temperature, lowest denaturant concentration, or lowest 'experimental parameter'), the amount of explained variance remains constant.

Case 1) Two basis spectra

The new set of basis spectra is calculated as

$$\begin{bmatrix} | & , & | \\ \phi_1 & , & \phi_2 \\ | & , & | \end{bmatrix} \cdot rotM = \begin{bmatrix} | & , & | \\ \phi_1' & , & \phi_2' \\ | & , & | \end{bmatrix} \qquad (41)$$

where ϕ_1 and ϕ_2 are the original first and second basis spectra, ϕ_1' and ϕ_2' are the new basis spectra, and $rotM$ is the rotation matrix given by

$$rotM = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \qquad (42)$$

where α is the angle of a right-triangle between a catheti of length $c_{1,1}$ and the hypotenuse of length $\sqrt{c_{1,1}^2 + c_{2,1}^2}$. $c_{1,1}$ is the associated coefficient for ϕ_1 and the first acquired spectrum. $c_{2,1}$ is the associated coefficient for ϕ_2 and the first acquired spectrum.

Case 2) Three basis spectra

The new set of basis spectra is calculated as

$$\begin{bmatrix} | & , & | & , & | \\ \phi_1 & , & \phi_2 & , & \phi_3 \\ | & , & | & , & | \end{bmatrix} \cdot rotZ \cdot rotY = \begin{bmatrix} | & , & | & , & | \\ \phi_1' & , & \phi_2' & , & \phi_3' \\ | & , & | & , & | \end{bmatrix} \quad (43)$$

where ϕ_1, ϕ_2 and ϕ_3 are the original first, second and third basis spectra, ϕ_1', ϕ_2' and ϕ_3' are the new basis spectra, $rotZ$ is the rotation matrix given by

$$rotZ = \begin{bmatrix} \cos(\gamma) & \sin(\gamma) & 0 \\ -\sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (44)$$

where γ is the angle of a right-triangle between a catheti of length $c_{1,1}$ and the hypotenuse of length $\sqrt{c_{1,1}^2 + c_{2,1}^2}$. Here, $c_{i,1}$ refers to the associated coefficient for i-th original basis spectrum and the first acquired spectrum. Then, $rotY$ is the rotation matrix given by

$$rotY = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \quad (45)$$

where β is the angle of a right-triangle between a catheti of length $\sqrt{c_{1,1}^2 + c_{2,1}^2}$ and the hypotenuse of length $\sqrt{c_{1,1}^2 + c_{2,1}^2 + c_{3,1}^2}$.

3.5. Alternative formulation for the explained variance

Given a set of orthogonal vectors (basis spectra) and the associated coefficients, the amount of explained variance for a given vector (basis spectrum) is calculated as:

$$\frac{||c_i||^2}{||\phi_i||^2} / ||M||^2 \qquad (46)$$

where ϕ_i is the i-th basis spectrum, c_i the associated coefficients, and M the data matrix. In the case of PCA, M is replaced by M_{meaned}.

4. Protein secondary structure estimation

Proteins containing different amounts of secondary structure elements, such as alpha helices, beta sheets, and random coils, produce different CD spectra[14]. Given a certain query spectrum and a reference set, the Selcon algorithm can be used to calculate the secondary structure content.[15]

4.1. The Selcon method

The Selcon algorithm can be divided into three versions (1, 2 and 3). ChiraKit will try to run Selcon 3 and if it fails it will run Selcon2.

Selcon version 1

We start with a set of linear equations that relate the CD spectra with the secondary structure content:

$$F = XC \qquad (47)$$

where C is the *m x n* matrix containing the CD spectra in delta epsilon units, F is the *l x n* matrix containing the known secondary structure elements, and X is the *l x m* matrix that allows the transition between the space of secondary structure elements F and the space of CD spectra C. Here, *n*, *l* and *m* represent, respectively, the number of proteins, secondary structure elements and wavelengths.

Let *c* be the query spectrum, an initial guess for the secondary structure components is done based on the most similar spectrum from the reference set, and we build the set of equations

---

[14] Greenfield, Norma J. "Using circular dichroism spectra to estimate protein secondary structure." Nature protocols 1.6 (2006): 2876-2890.

[15] Sreerama, Narasimha, Sergei Yu Venyaminov, and Robert W. Woody. "Estimation of protein secondary structure from circular dichroism spectra: inclusion of denatured proteins with native proteins in the analysis." Analytical biochemistry 287.2 (2000): 243-251.

$$F_2 = XC_2 \quad (48)$$

where C_2 now includes c, has $(n + 1)$ columns, and is ordered according to the root mean square (RMS) distance to c. F2 is also rearranged accordingly. The equation is solved for X and then the secondary structure f (of the protein to be analysed) is calculated as

$$f = Xc \quad (49)$$

To solve the equation regarding F_2, C_2 is factorized using the SVD algorithm. Indeed, by varying the number of proteins from the reference set (from 2 to n) and the number of relevant singular values (from 1 to 7), several solutions are obtained for f. Those satisfying the following criteria are considered: 1) the sum of the secondary structure elements' fractions lies between 95% and 105%, and 2) each secondary structure element fraction is greater than - 2.5 %. If no acceptable solutions are found during this step, the rules are relaxed by adding a tolerance of 1 % to the total sum of fractions, and of 0.5 % to the individual fractions.

For each set of valid solutions that share a certain number of proteins from the reference set but vary in the number of relevant singular values, only the solution whose sum of secondary structure element fractions is closest to 100% is kept. Let N_p be the number of valid solutions, we will have at most $2 \leq N_p \leq n$ solutions.

The N_p solutions are averaged to obtain f_1 , which replaces the initial guess in the matrix F_2 as the second approximation. Finally, the same process is repeated until the RMS difference between two successive solutions is less than 0.0025 delta epsilon units.

Selcon version 2

The Selcon Version 2 method starts with the same algorithm as Selcon Version 1. After self-consistency is achieved, the final N_p set of valid solutions are filtered by removing those where the RMS between the reconstructed spectrum and the query spectrum c is larger than 0.25 delta epsilon units. If the filter does not leave any solution, the tolerance is relaxed using 0.01 delta epsilon units steps until at least one solution is found. Finally, the mean solution of the filtered solutions is reported.

Selcon version 3

The Selcon Version 3 method follows the same procedure as Selcon Version 2 with the difference that one more rule is applied. The N_p set of valid solutions is further filtered based on the amount of alpha helix.

Let h_i be the alpha-helix content fraction of the i-th solution, h_q the predicted alpha-helix content fraction for the query spectrum (after the Selcon 2 method); and h_{min} , h_{ave} and h_{max} the minimum, average, and maximum fraction of alpha-helix content, respectively, considering all the reference proteins required to obtain the N_p solutions. Then, for the i-th solution to be valid, it needs to satisfy any of these criteria:

Case 1) $$h_q > 0.65 \;\&\; h_i > 0.65 \quad (50)$$

Case 2) $$0.25 \le h_q \le 0.65 \;\&\; \frac{h_q + h_{max}}{2} - 0.03 \le h_i \le \frac{h_q + h_{max}}{2} + 0.03 \quad (51)$$

Case 3) $$0.15 \le h_q \le 0.25 \;\&\; \frac{h_q + h_{ave}}{2} - 0.03 \le h_i \le \frac{h_q + h_{ave}}{2} + 0.03 \quad (52)$$

Case 4) $$h_q < 0.15 \;\&\; \frac{h_q + h_{min}}{2} - 0.03 \le h_i \le \frac{h_q + h_{min}}{2} + 0.03 \quad (53)$$

4.2. Reference sets

By default, two reference sets are provided: AU_SP175 and 2) AU_SM180. These sets contain CD spectra from the same group of proteins as the original SP175[16] and SMP180[17] sets. AU_SP175 contains the CD spectra of 71 soluble proteins measured down to 175 nm. AU_SM180 contains the AU_SP175 set and 57 additional CD spectra, including membrane proteins. The lower limit of the AU_SM180 set is 180 nm. In both cases, the data sets have 1 nm steps and start at 240 nm.

The number of secondary structure elements to be detected depends on the reference set (or the lower wavelength limit). Six secondary structure elements for values lower than 180 nm (AU_SP175), and four secondary structure elements for values between 180 and 190 nm (AU_SM180). **No calculation is possible if the query spectrum does not go to at least 190 nm.**

To select the reference data, the CD spectra were downloaded from the Protein Circular Dichroism Data Bank (PCDDB)[18]. In case that many versions were available, the latest one was chosen (September 2023). In two cases we did not use the PCDDB data directly: The Human Serum Albumin CD spectrum was replaced

---

[16] Lees, Jonathan G., et al. "A reference database for circular dichroism spectroscopy covering fold and secondary structure space." Bioinformatics 22.16 (2006): 1955-1962.

[17] Abdul-Gader, Ali, Andrew John Miles, and Bonnie A. Wallace. "A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy." Bioinformatics 27.12 (2011): 1630-1636.
[18] https://pcddb.cryst.bbk.ac.uk/

with an in-house measured spectrum. The Lysozyme spectrum CD0000045100 was zeroed again and averaged with the spectrum CD0000045000.

Custom reference sets can be imported too. The requisite is to import the corresponding matrices F (secondary structure elements) and C (reference spectra). C is arranged column-wise (one spectrum per column) and the CD data goes from the highest wavelength to the lowest wavelength. F should have the same columns as C, and as many rows as different secondary structure elements.

4.3. Secondary structure elements

All secondary structures elements come from the DSSP routine[19]. Following the method of Sreerama, Narasimha et al., 1999, the secondary structure elements are further categorised as regular or distorted alpha-helix, regular or distorted beta-sheet, turns and 'other'. Alpha helix are the sum of fractions H and G. Beta sheets are the fraction E. Turns are fraction T. All other fractions (I, B, S, C) are assigned to ‘other’. The alpha helix fraction is divided into two groups: Alpha_D (distorted) which are the two first and two last residues in a helix and the rest are Alpha_R (regular). If the helix is less than 4 residues all go into the Alpha_D fraction. The same division is made for beta sheets, with the difference that only the first and the last residue of a beta strand goes into the distorted fraction Beta_D. The rest goes into the regular fraction Beta_R.

For a given spectrum with an unknown associated secondary structure, if the lower wavelength range value is below 180 nm, the six secondary structure elements can be detected. If the lower limit is between 180 and 190 nm (inclusive), four different fractions will be estimated: Alpha, Beta, Turns and Other (Alpha = Alpha_R + Alpha_D, Beta = Beta_R + Beta_D).

5. Peptide helicity estimation

Certain peptides adopt only helix or coil conformations. The helix conformation has two minima at 208 and 222 nm, while the coil conformation has a negative peak at 200 nm. If the peptide can only populate these two states, an isodichroic point around 203 nm should be observed for the transition[20]. The minima at 222 nm can be used to estimate the peptide helicity.[21]

---

[19] Kabsch, Wolfgang, and Christian Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen‐bonded and geometrical features." Biopolymers: Original Research on Biomolecules 22.12 (1983): 2577-2637.
[20] Toniolo, Claudio, Fernando Formaggio, and Robert W. Woody. "Electronic circular dichroism of peptides." Comprehensive chiroptical spectroscopy: applications in stereochemical analysis of synthetic compounds, natural products, and biomolecules 2 (2012): 499-544.
[21] Zavrtanik, Uroš, Jurij Lah, and San Hadži. "Estimation of Peptide Helicity from Circular Dichroism Using the Ensemble Model." The Journal of Physical Chemistry B 128.11 (2024): 2652-2663.

5.1. The ensemble model

The ensemble model assumes that there exists different helical conformers of varying lengths, and that each of them produces different signals. The peptide can have one or two helical segments. Triple or higher helical segments are not considered because of their low probability. The model is explained in Uroš Zavrtanik *et. al*, 2024 and implemented in https://github.com/sanhadzi/Dichroic-CD-model.

Contact details

For further assistance, please contact us:

📧 spc@embl-hamburg.de
📌EMBL (c/o DESY), Notkestrasse 85, Build. 25a, 22607 Hamburg, Germany