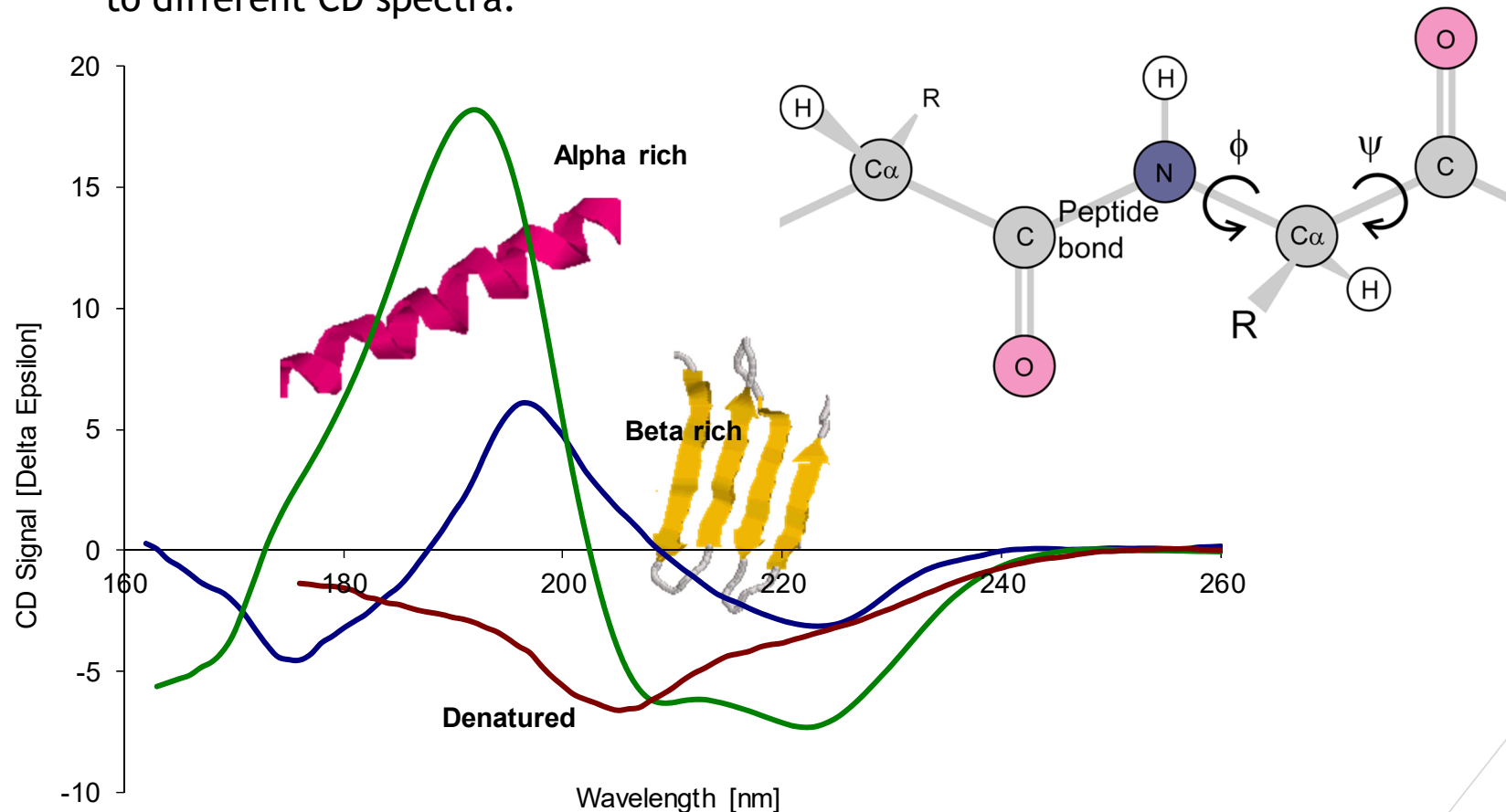


# ESC1: Circular Dichroism: best practice and data analysis

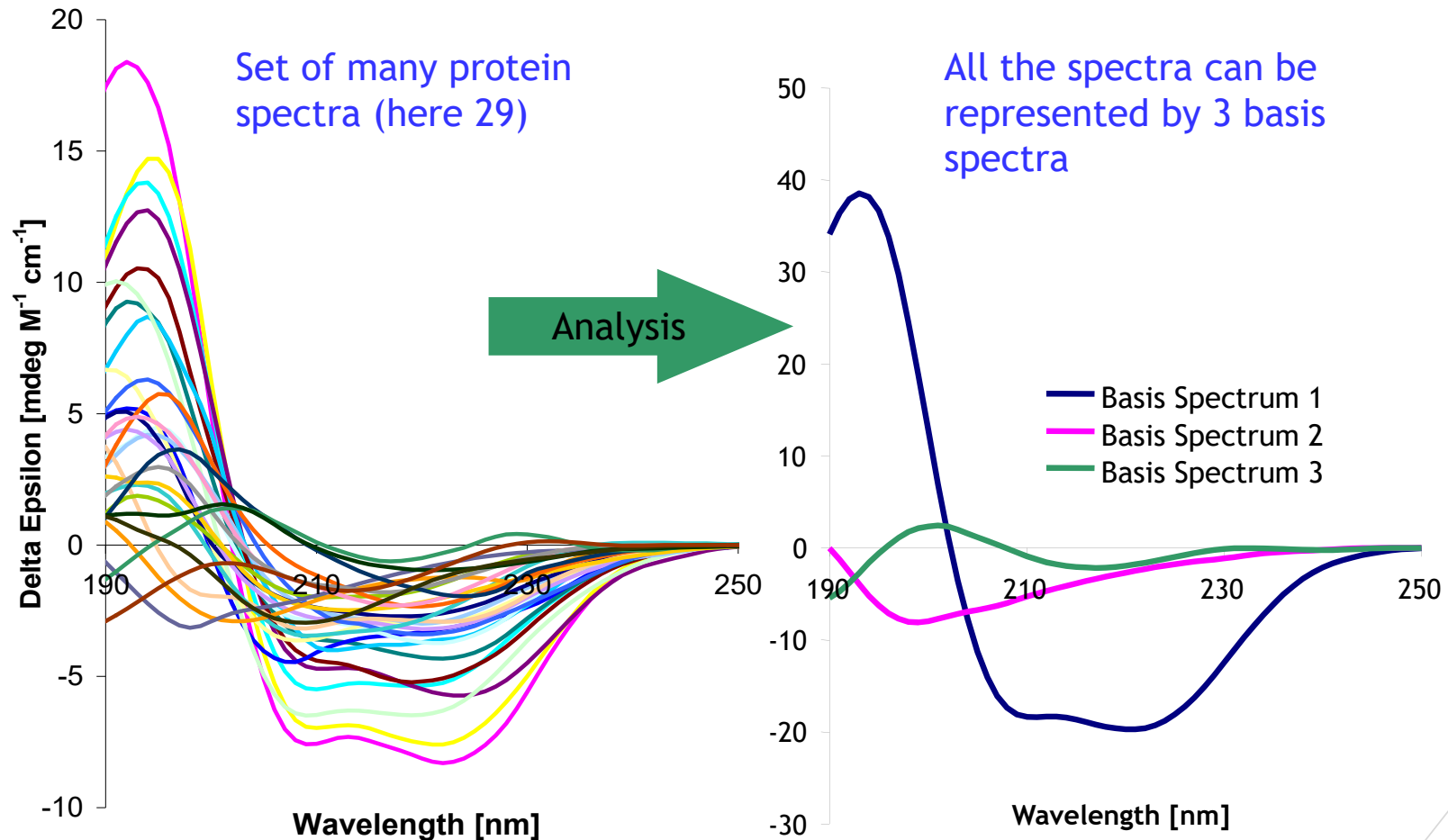
**Lecture 4:** Secondary structure calculations. How to do it and understand the limits

# Information in a protein CD spectrum

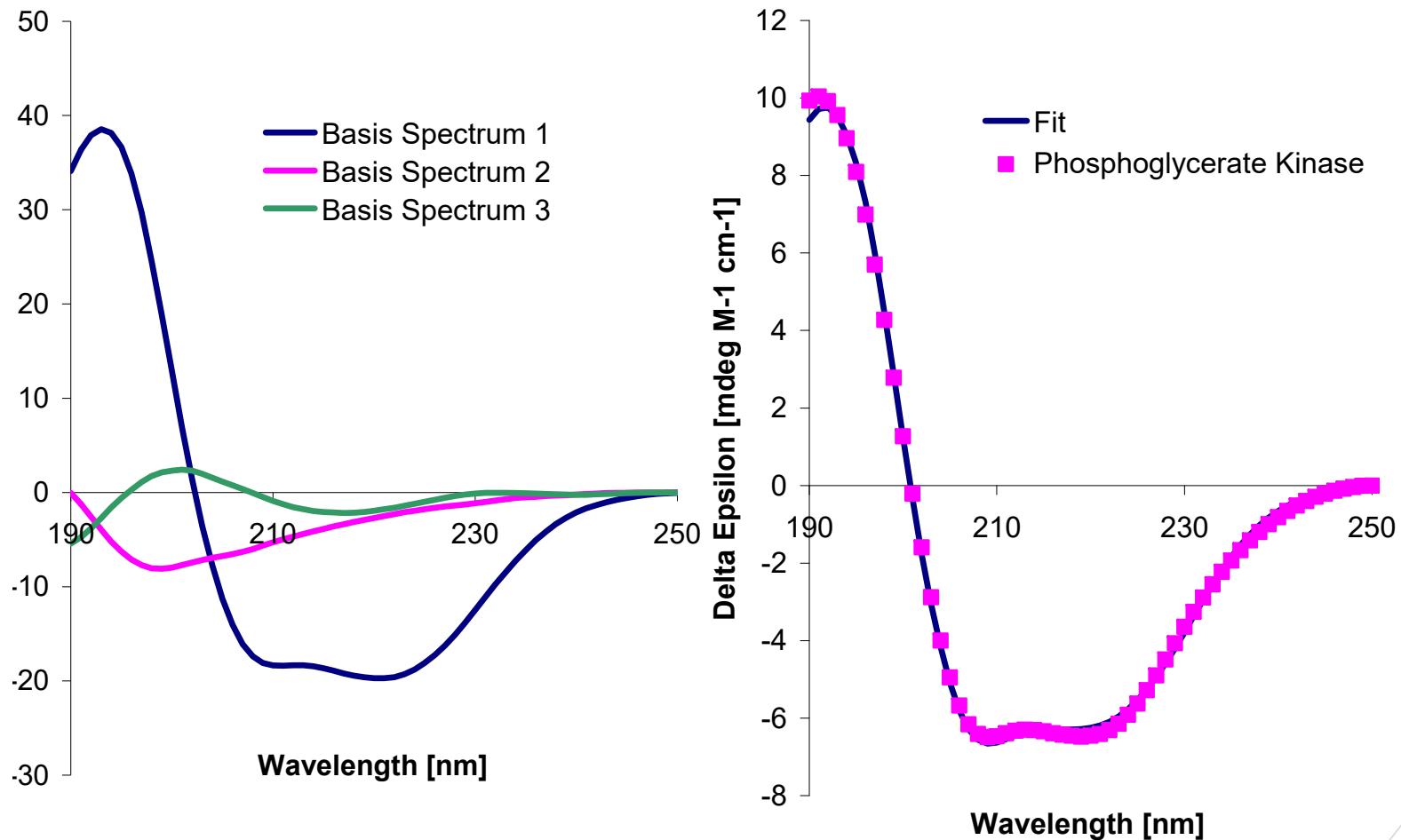
Different secondary structure in proteins gives rise to different CD spectra.



# Information in a protein CD spectrum



# All spectra may be reconstructed from the three basis spectra



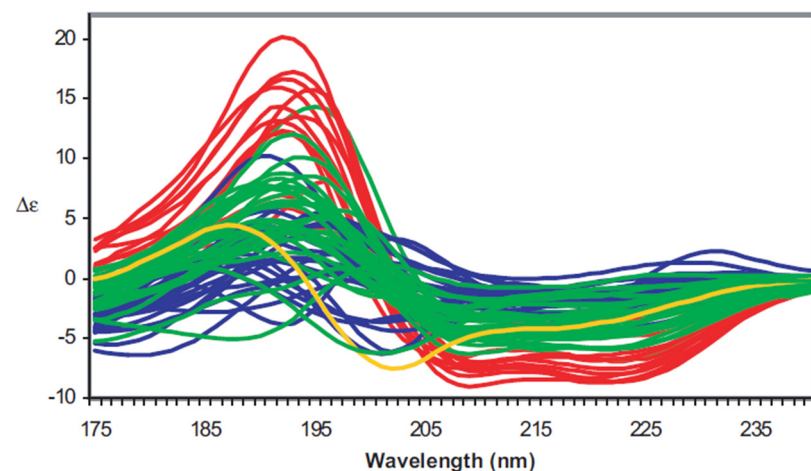
# CD spectrum analysis

## Questions:

- What is the analysis which decomposes a set of known CD spectra into their basis spectra?
- How is the Secondary Structure calculated from the set of known CD Spectra?

## What do we have:

71 CD spectra of proteins



All the protein secondary structures  
(from crystallography)

	Aldolase	Alkaline phosphatase	Alpha amylase	.....
$\alpha$ -helix	0.458	0.311	0.285	.....
$\beta$ -sheet	0.138	0.183	0.205	.....
Turns	0.103	0.133	0.122	.....
Other	0.301	0.374	0.388	.....

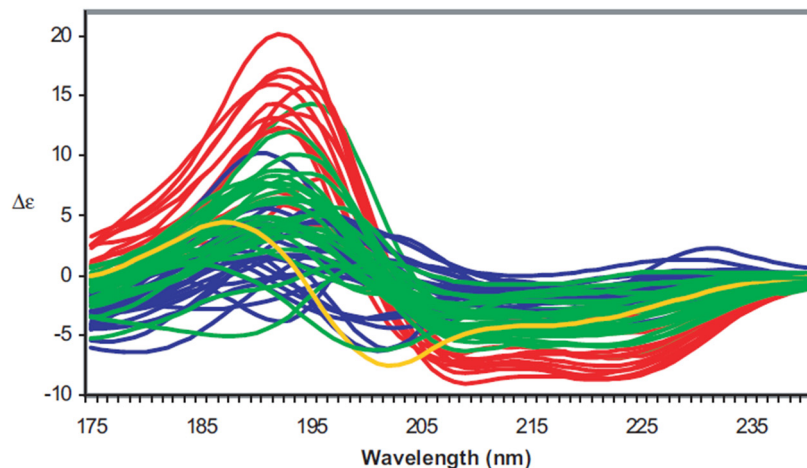


*Or any other classification!*

SP175 reference dataset: J.G. Lees *et al.* Bioinformatics. 22 (2006) 1955-1962

# CD spectrum analysis

71 CD spectra of proteins



Write them as Matrices

$$A_{\text{ref}} = \begin{bmatrix} \text{CD of Aldolase} \\ \text{CD of Alkaline...} \\ \text{CD of Alpha...} \\ \dots \\ \text{CD of Ubiquitin} \end{bmatrix}$$

**m x 71 matrix**

where m is number of data points

All the protein secondary structures  
(from crystallography)

	Aldolase	Alkaline phosphatase	Alpha amylase	.....
$\alpha$ -helix	0.458	0.311	0.285	.....
$\beta$ -sheet	0.138	0.183	0.205	.....
Turns	0.103	0.133	0.122	.....
Other	0.301	0.374	0.388	.....



*Or any other classification!*

$$F_{\text{ref}} = \begin{bmatrix} \text{Aldolase} & \text{Alkaline...} & \dots & \text{Ubiquitin} \end{bmatrix} \begin{bmatrix} 0.458 & 0.311 & \dots & 0.250 \\ 0.138 & 0.183 & \dots & 0.316 \\ 0.103 & 0.133 & \dots & 0.145 \\ 0.301 & 0.374 & \dots & 0.289 \end{bmatrix} \begin{matrix} \alpha\text{-helix } (\alpha) \\ \beta\text{-sheet } (\beta) \\ \text{Turns } (T) \\ \text{Other } (O) \end{matrix}$$

**l x 71 matrix**

where l is number of  
secondary structures  
(here 4, often 6)

# CD spectrum analysis

$$A_{\text{ref}} = \begin{bmatrix} \text{CD of Aldolase} \\ \text{CD of Alkaline...} \\ \text{CD of Alpha...} \\ \dots \\ \text{CD of Ubiquitin} \end{bmatrix}$$

**m x 71 matrix**

where m is number of data points

$$F_{\text{ref}} = \begin{bmatrix} \text{Aldolase} & \text{Alkaline...} & & \text{Ubiquitin} \\ 0.458 & 0.311 & & 0.250 \\ 0.138 & 0.183 & & 0.316 \\ 0.103 & 0.133 & \dots\dots\dots & 0.145 \\ 0.301 & 0.374 & & 0.289 \end{bmatrix} \begin{matrix} \alpha\text{-helix } (\alpha) \\ \beta\text{-sheet } (\beta) \\ \text{Turns } (T) \\ \text{Other } (O) \end{matrix}$$

**l x 71 matrix**

where l is number of secondary structures

Does a matrix X (l x m) exists such that? :

$$F = X A$$

*Note that A is not a square matrix, so there is no inverse  $A^{-1}$*

- The Singular Value Decomposition (SVD) theorem:

U, S and V exists:  $A = USV^T$  where  $V^T V = U U^T = I$ ,  $S^+$  exists  $SS^+ = I$

- This means that we can calculate X

$$F = X A = X USV^T \Rightarrow X = FVS^+U^T$$

- Basis spectra: US    Coefficients which recreates the reference spectra:  $V^T$

**X exists via Singular Value Decomposition (SVD)**

Short cut



# CD spectrum analysis

$$A_{\text{ref}} = \begin{bmatrix} \text{CD of Aldolase} \\ \text{CD of Alkaline...} \\ \text{CD of Alpha...} \\ \dots \\ \text{CD of Ubiquitin} \end{bmatrix}$$

$$F_{\text{ref}} = \begin{bmatrix} \text{Aldolase} & \text{Alkaline...} & & \text{Ubiquitin} \\ 0.458 & 0.311 & & 0.250 \\ 0.138 & 0.183 & & 0.316 \\ 0.103 & 0.133 & \dots\dots\dots & 0.145 \\ 0.301 & 0.374 & & 0.289 \end{bmatrix} \begin{matrix} \alpha\text{-helix } (\alpha) \\ \beta\text{-sheet } (\beta) \\ \text{Turns } (T) \\ \text{Other } (O) \end{matrix}$$

$F = X A$  *X exists via Singular Value Decomposition (SVD) and can be calculated*

*In principle*, for a CD spectrum  $C = \begin{bmatrix} \text{CD of protein} \end{bmatrix}$  of a protein of unknown structure :  $f = \begin{bmatrix} \alpha \\ \beta \\ T \\ O \end{bmatrix}$

Then  $\begin{bmatrix} \alpha \\ \beta \\ T \\ O \end{bmatrix} = X \begin{bmatrix} \text{CD of protein} \end{bmatrix}$  or  $f = XC$  can be calculated

*In practice*, this seldom gives a valid solution!

- The sum  $\alpha + \beta + T + O$  is in general not 100%
- Each component may even be  $< 0$



# CD spectrum analysis

$$f = \begin{bmatrix} \alpha \\ \beta \\ T \\ O \end{bmatrix} \text{ from } A_{\text{ref}} = \begin{bmatrix} \text{CD of Aldolase} & \text{CD of Alkaline...} & \text{CD of Alpha...} & \dots & \text{CD of Ubiquitin} \end{bmatrix} C = \begin{bmatrix} \text{CD of protein} \end{bmatrix} F_{\text{ref}} = \begin{bmatrix} \text{Aldolase} & \text{Alkaline...} & \dots & \text{Ubiquitin} \\ 0.458 & 0.311 & & 0.250 \\ 0.138 & 0.183 & & 0.316 \\ 0.103 & 0.133 & & 0.145 \\ 0.301 & 0.374 & & 0.289 \end{bmatrix}$$

Several methods exist which use SVD and finds the structure  $f = \begin{bmatrix} \alpha \\ \beta \\ T \\ O \end{bmatrix}$  based on A and F from C

We often use and do recommend these methods

- **Selcon3**: Self-consistent method (next slide)
- **CDSSTR**: Find f via SVD from a random selection of 8 spectra from  $A_{\text{ref}}$ .
- **CONTIN/LL**: Fit the spectrum C to a combination of spectra in  $A_{\text{ref}}$  resembling C the most

# CD spectrum analysis

$$f = \begin{bmatrix} \alpha \\ \beta \\ T \\ 0 \end{bmatrix} \text{ from } A_{\text{ref}} = \begin{bmatrix} \text{CD of Aldolase} \\ \text{CD of Alkaline...} \\ \text{CD of Alpha...} \\ \dots \\ \text{CD of Ubiquitin} \end{bmatrix} C = \begin{bmatrix} \text{CD of protein} \end{bmatrix} F_{\text{ref}} = \begin{bmatrix} \text{Aldolase} & \text{Alkaline...} & \dots & \text{Ubiquitin} \\ 0.458 & 0.311 & & 0.250 \\ 0.138 & 0.183 & & 0.316 \\ 0.103 & 0.133 & & 0.145 \\ 0.301 & 0.374 & & 0.289 \end{bmatrix}$$

As an example, Selcon3 does:

- Sort  $A_{\text{ref}}$  such that the spectrum which resembles  $C$  the most are to the left
- Guess a solution  $f_{\text{guess}}$  (from the spectrum resembling  $C$  the most)
- Add  $C$  to  $A_{\text{sort}}$  and  $f_{\text{guess}}$  to  $F_{\text{ref}}$ :  $A = [C, A_{\text{sort}}]$  and  $F = [f_{\text{guess}}, F_{\text{sort}}]$
- Use SVD to find solutions using from 3 and up of the spectra in the new  $A$
- Collect all solutions and select the valid ones:
  - Sum Structures = 100%+-5% and each structure > -2.5%
- Use the average of the  $f$  solutions to make a new  $f_{\text{guess}}$

Repeat the until  
self-consistency

Repeat the until *self-consistency*, i.e. new  $f$  solution is close to previous  $f_{\text{guess}}$

# CD spectrum analysis

## Selcon3

$$1) \text{ Sort and guess } A_{\text{sort}} = \begin{bmatrix} \text{CD Resemble C most} \\ \text{CD Resemble C less} \\ \dots \\ \text{CD Resemble C least} \end{bmatrix}, F_{\text{sort}} = \begin{bmatrix} f \text{ Resemble C most} \\ f \text{ Resemble C less} \\ \dots \\ f \text{ Resemble C least} \end{bmatrix}, f_{\text{guess}} = \begin{bmatrix} \alpha_{\text{g}} \\ \beta_{\text{g}} \\ \tau_{\text{g}} \\ \text{O}_{\text{g}} \end{bmatrix} = \begin{bmatrix} f \text{ Resemble C most} \end{bmatrix}$$

$$2) \text{ Do SVD on several subsets of } A = [C A_{\text{sort}}] \text{ and } F = [f_{\text{guess}} F_{\text{sort}}]$$

3) Find all valid solutions  $f_{\text{valid},i}$  :  $\alpha + \beta + \tau + \text{O} \sim 100\% \pm 5\%$  and  $\alpha, \beta, \tau, \text{O} > -2.5\%$

4) Use the new solution  $f_{\text{sol}} = \text{average}(f_{\text{valid},i})$  as a new  $f_{\text{guess}}$

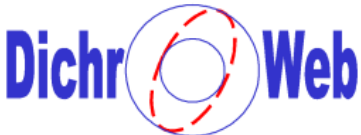
5) *Repeat* 2 - 4 until  $f_{\text{sol}} \sim f_{\text{guess}}$

By repeating, Selcon3 ensures *self-consistency* in solutions

# CD spectrum analysis

← → ↻ ⓘ dichroweb.cryst.bbk.ac.uk/html/home.shtml

[Home](#)  
[Input Data](#)  
[User Guide](#)  
[Background Information](#)  
[FAQ](#)  
[References](#)  
[Links](#)  
[Contact Us](#)  
[Terms and Conditions](#)  
[Cookies](#)



## On-line analysis for protein Circular Dichroism spectra

[Apply for a user-account](#)

[Analyse data](#) (registered users only)

**Citing DichroWeb:**  
If you use DichroWeb for your analysis you agree to cite the publications detailing the original methods and reference data used, as well as one of the specific DichroWeb papers:

**Whitmore, L. and Wallace, B.A. (2008) Biopolymers 89: 392-400. (PDF)**

**Whitmore, L. and Wallace, B.A. (2004) Nucleic Acids Research 32: W668-673. (PDF)**

### DichroWeb News

Analyses now possible using Membrane Protein data set SMP180. Abdul-Gader A, Miles AJ, Wallace BA. Bioinformatics (2011) 27 1630-6.

Video guides:

- ★ [\[new\] Accurate measuring of the true pathlength of optical CD cells](#)
- ★ [Cleaning and Loading Circular Dichroism Cells](#)
- ★ [Calibrating CD Spectra with CDTool and MS Excel](#)
- ★ [Measuring a CSA spectrum](#)
- ★ [PCDDDB Tutorial](#)
- ★ [Analysing Protein CD Data using Dichroweb](#)

Related Projects [ValiDichro: CD validation and quality control](#), [2Struc: The Secondary Structure Server](#), [Dichromatch](#), and the [Protein Circular Dichroism Data Bank](#) are now open for use.

### Stats

DichroWeb currently has 5900+ registered users and has performed over 680,000 deconvolutions.

DichroWeb server: A.J. Miles *et al.* Protein Science. 2021 doi: 10.1002/pro.4153



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004806

CD BOOK (contains tips on successful DichroWeb use) : Modern Techniques for Circular Dichroism and Synchrotron Radiation Circular Dichroism Spectroscopy; Eds: Wallace and Janes. [Click for Info](#)

## Registration:

UserID:  [Help](#)

IDpassword:  [Help](#)

Log-in

## Input File Details:

Protein name:  [Help](#)

File location:  No file chosen [Help](#)

File with CD spectrum

## About the Data File:

File Format:  [Help](#)

Input Units:  [Help](#)

Initial Wavelength (nm):  [Help](#)

Final Wavelength (nm):  [Help](#)

Units and wavelength range

NB: The initial wavelength should match the wavelength nearest the top of your data file, as you read it.

Wavelength Step (nm): ☐ 0.05 ☐ 0.1 ☐ 0.2 ☐ 0.5 ☒ 1.0 [Help](#)

Lowest nm datapoint to use in the analysis:  [Help](#)

## Choice of Methods:

Analysis Programme:  [Help](#)

Reference Set:  [Help](#)

## Advanced Options:

Optional Scaling Factor:  [Help](#)

## Output Options:

Output Units:  [Help](#)

SERVICE INFORMATION

submit

agreemer

SELCON3  
CONTIN  
~~VARSLC (no reference set required)~~  
CDSSTR  
~~K2D (no reference set required)~~

~~Set 1 (Optimised for 178-260 nm)  
Set 2 (Optimised for 178-260 nm)  
Set 3 (Optimised for 185-240 nm)  
Set 4 (Optimised for 190-240 nm)  
Set 5 (Optimised for 178-260 nm)  
Set 6 (Optimised for 185-240 nm)  
Set 7 (Optimised for 190-240 nm)  
SP175 (Optimised for 175-240 nm # Not for Selcon3.)  
SP175 (Optimised for 190-240 nm # Less nm required.)  
Cryst175 (Optimised for 175-240 nm # Not for Selcon3.)  
SMP180 (Optimised for 180-240 nm)  
SMP180 (Optimised for 190-240 nm # Less nm required.)~~

[Home](#)

[Input Data](#)

[User Guide](#)

[Background Information](#)

[FAQ](#)

[References](#)

[Links](#)

[Contact Us](#)

[Terms and Conditions](#)

[Cookies](#)

# CD spectrum analysis

## Choice of Methods:

Analysis Programme:  [Help](#)

Reference Set:  [Help](#)

## Reference CD data:

*Collection of (SR)CD data of proteins with known secondary structure*

**SP175:** Soluble proteins

**SMP180:** Membrane proteins

Data down to 175 and 180 nm, respectively.

SP175 (Optimised for 175-240 nm # Not for Selcon3.)  
SP175 (Optimised for 190-240 nm # Less nm required.)  
Cryst175 (Optimised for 175-240 nm # Not for Selcon3.)  
SMP180 (Optimised for 180-240 nm)  
SMP180 (Optimised for 190-240 nm # Less nm required)

## Three different (mathematical) methods:

*Sort reference protein CD data and compare to the measured CD spectrum.*

Please use all three methods and compare results.

Two of the methods should give comparable amounts of  $\alpha$ -helix,  $\beta$ -sheet etc.

SELCON3  
CONTIN  
~~VARSLC (no reference set required)~~  
CDSSTR  
~~ICD (no reference set required)~~



# CD spectrum analysis: Example

Human osteopontin (hOPN)

hOPN

Solutions from the CDSSTR method

Solutions using reference database: SP175.

Use of the reference set requires the citation of:

Lees, J.G., Miles, A.J., Wien, F., and Wallace, B.A. (2006), Bioinformatics, 22, 1955-1962.

NRMSD:0.038

Helix segments per 100 residues: 1.708

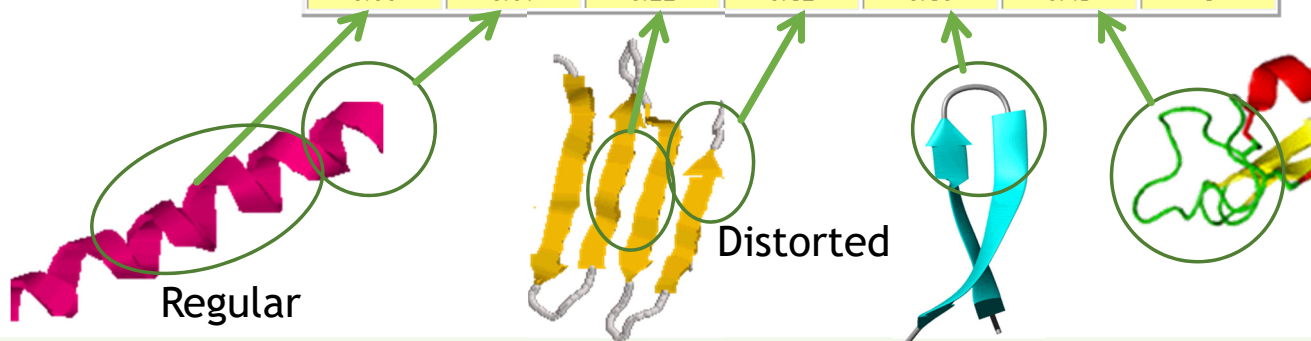
Ave helix length per segment: 3.769

Strand segments per 100 residues: 6.104

Ave strand length per segment: 5.536

Calculated Secondary Structure Fractions

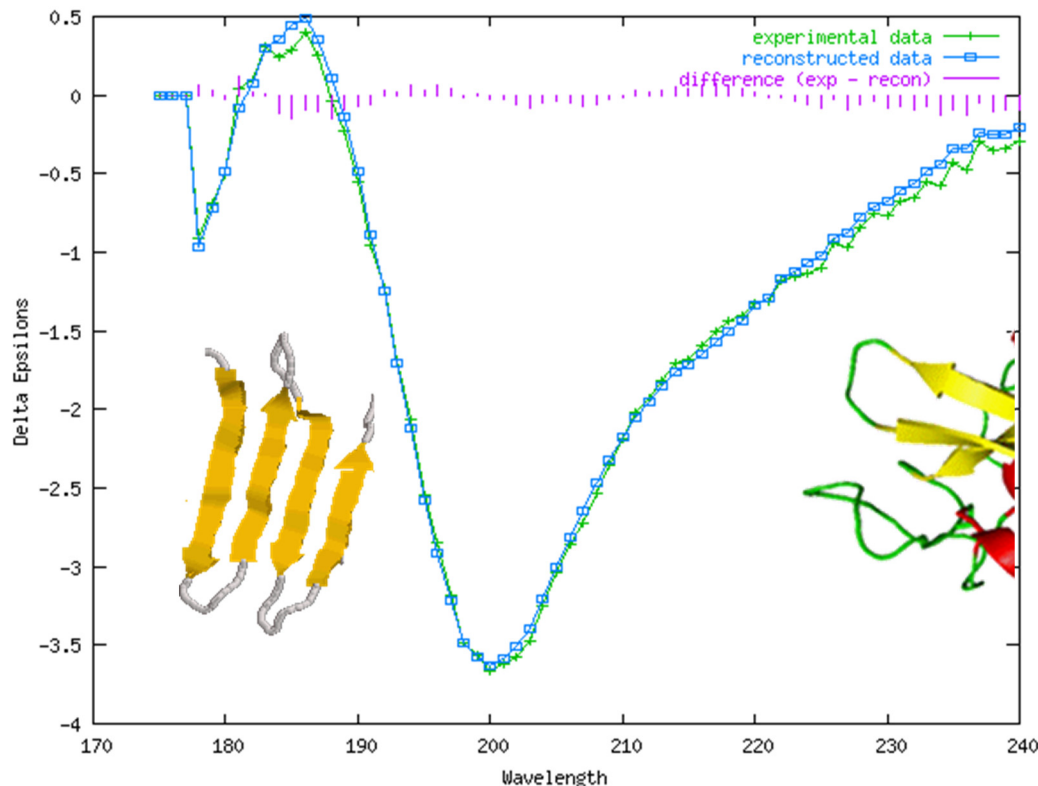
Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
0.00	0.07	0.22	0.12	0.16	0.43	1



# CD spectrum analysis: Example

Human osteopontin (hOPN)

Dichroweb also displays a calculated CD spectrum for comparison.



Typical spectrum for a highly unordered protein.



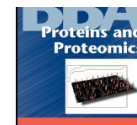


# CD spectrum analysis: Example



Biochimica et Biophysica Acta

journal homepage: [www.elsevier.com/locate/bbapap](http://www.elsevier.com/locate/bbapap)



## Multiple low-affinity interactions support binding of human osteopontin to integrin $\alpha_x\beta_2$



Eva Klänning<sup>a,b</sup>, Brian Christensen<sup>a</sup>, Goran Bajic<sup>a</sup>, Søren V. Hoffmann<sup>c</sup>, Nykola C. Jones<sup>c</sup>, Morten M. Callesen<sup>a</sup>, Gregers R. Andersen<sup>a</sup>, Esben S. Sørensen<sup>a,d</sup>, Thomas Vorup-Jensen<sup>b,d,e,\*</sup>

<sup>a</sup> Dept. of Molecular Biology and Genetics Aarhus University, Aarhus, Denmark

<sup>b</sup> Dept. of Biomedicine, Denmark

<sup>c</sup> Institute for Storage Ring Facilities Aarhus (ISA), Dept. of Physics and Astronomy & Center for Storage Ring Facilities Aarhus, Denmark

<sup>d</sup> Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Aarhus Denmark

<sup>e</sup> MEMBRANES Research Center, Aarhus University, Aarhus, Denmark

Biochimica et Biophysica Acta 1854 (2015) 930–938

**Table 2**

Secondary structures of human OPN, dOPN and HCM.

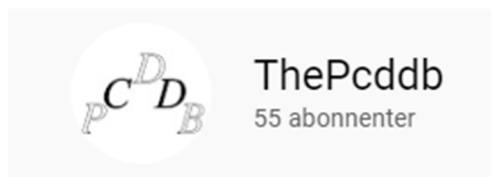
The content of regular  $\alpha$ -helical<sup>a</sup>, disordered  $\alpha$ -helix<sup>b</sup>, regular  $\beta$ -strand<sup>c</sup>, disordered  $\beta$ -strand<sup>d</sup>, turns<sup>e</sup> and unordered conformation<sup>f</sup> in OPN, dOPN, HCM and  $\beta$ -casein. Results for OPN and dOPN are averaged from at least three independent experiments performed in triplicate whereas results for HCM represent single experiment where measurements were performed in triplicate. The measurements for calculating the distribution of secondary structure were collected at 25 °C.

Protein	Regular $\alpha$ -helix <sup>a</sup>	Distorted $\alpha$ -helix <sup>b</sup>	Regular $\beta$ -strand <sup>c</sup>	Distorted $\beta$ -strand <sup>d</sup>	Turns <sup>e</sup>	Unordered <sup>f</sup>	Total
OPN	0.00	0.06	0.22	0.13	0.14	0.44	0.99
dOPN	0.00	0.06	0.22	0.13	0.14	0.44	0.99
HCM	0.10	0.12	0.11	0.08	0.13	0.46	1



# CD spectrum analysis: DichroWeb

If you like video tutorials,  
have a look at the  
YouTube channel



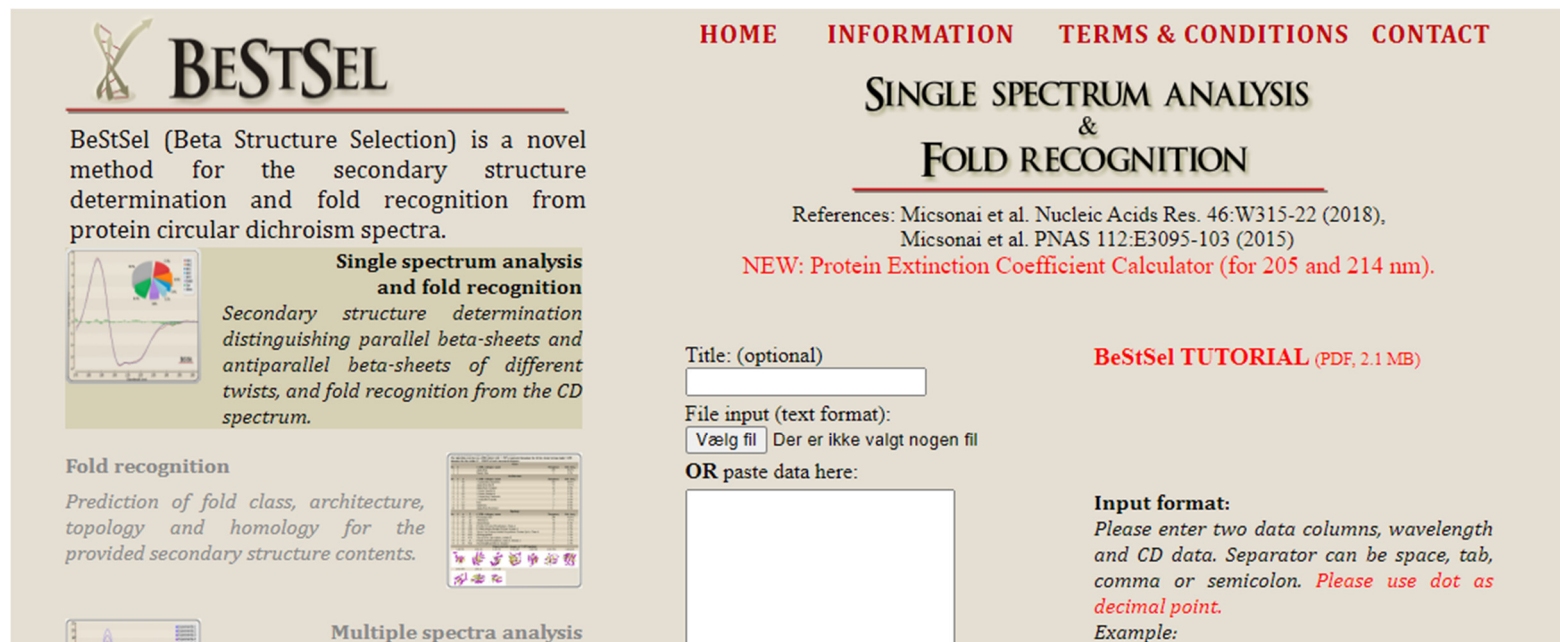
[youtube.com/user/ThePcddb/videos](https://youtube.com/user/ThePcddb/videos)

A screenshot of the DichroWeb job submission form in a Mozilla Firefox browser window. The browser's address bar shows the URL 'dichroweb.cryst.bbk.ac.uk/html/process.shtml'. The page has a blue sidebar with navigation links: Home, Input Data, User Guide, Background Information, FAQ, References, Links, Contact Us, Terms and Conditions, and Cookies. The main content area features the DichroWeb logo, a 'Server Status' box indicating 'No reported problems', and a link to a 'CD BOOK'. Below this is a 'Registration' section with fields for 'UserID' (gmile01) and 'IDpassword'. The 'Input File Details' section includes fields for 'Protein name' (ms13) and 'File location' (ss09266.gen), with a note about using ASCII files. The 'About the Data File' section has dropdowns for 'File Format' (FREE) and 'Input Units' (delta epsilon), and input fields for 'Initial Wavelength (nm)', 'Final Wavelength (nm)', and 'Wavelength Step (nm)' (0.1, 0.2, 0.5, 1.0). A 'Lowest nm datapoint to use in the analysis' field is at the bottom. The browser's status bar shows '1:29 / 6:05'.

# CD spectrum analysis: BestSel

Another good option for secondary structure analysis is BestSel

<https://bestsel.elte.hu/index.php>



The screenshot shows the BestSel web interface. At the top, there is a navigation bar with links: HOME, INFORMATION, TERMS & CONDITIONS, and CONTACT. The main heading is "SINGLE SPECTRUM ANALYSIS & FOLD RECOGNITION". Below this, references are listed: Micsonai et al. Nucleic Acids Res. 46:W315-22 (2018), and Micsonai et al. PNAS 112:E3095-103 (2015). A new feature is highlighted: "NEW: Protein Extinction Coefficient Calculator (for 205 and 214 nm)". The interface is divided into several sections: "Single spectrum analysis and fold recognition" with a description of the method and a graph; "Fold recognition" with a description of the prediction capabilities and a screenshot of the output; and "Multiple spectra analysis" with a small icon. On the right side, there is a "BeStSel TUTORIAL (PDF, 2.1 MB)" link. Below the navigation bar, there are input fields for "Title (optional)", "File input (text format)" (with a "Vælg fil" button and a message "Der er ikke valgt nogen fil"), and "OR paste data here:". An "Input format:" section provides instructions: "Please enter two data columns, wavelength and CD data. Separator can be space, tab, comma or semicolon. Please use dot as decimal point." and an "Example:".

*BestSel has more emphasis on different beta-sheet structures*

# CD spectrum analysis: Good practice and limits

In the SP175 reference set, great care was taken to determine the concentrations

- Do the best to get a *good concentration measurement*. Within 5-10%
- You may scale your spectrum a bit up/down and check the results but large scale factors are not recommended
- Determine concentrations using multiple methods if absolute values are crucial.

*Absolute secondary structure values are less reliable than relative values*

- Absolute values should be taken with care. Use several methods and check how reliable/comparable they are
- If you have a series of CD spectra on the same sample under different conditions, the relative changes are generally more reliable
- Do not quote absolute values with many digits, even if the analysis program presents them to you

