MOSBRI
Molecular-Scale Biophysics Research Infrastructure

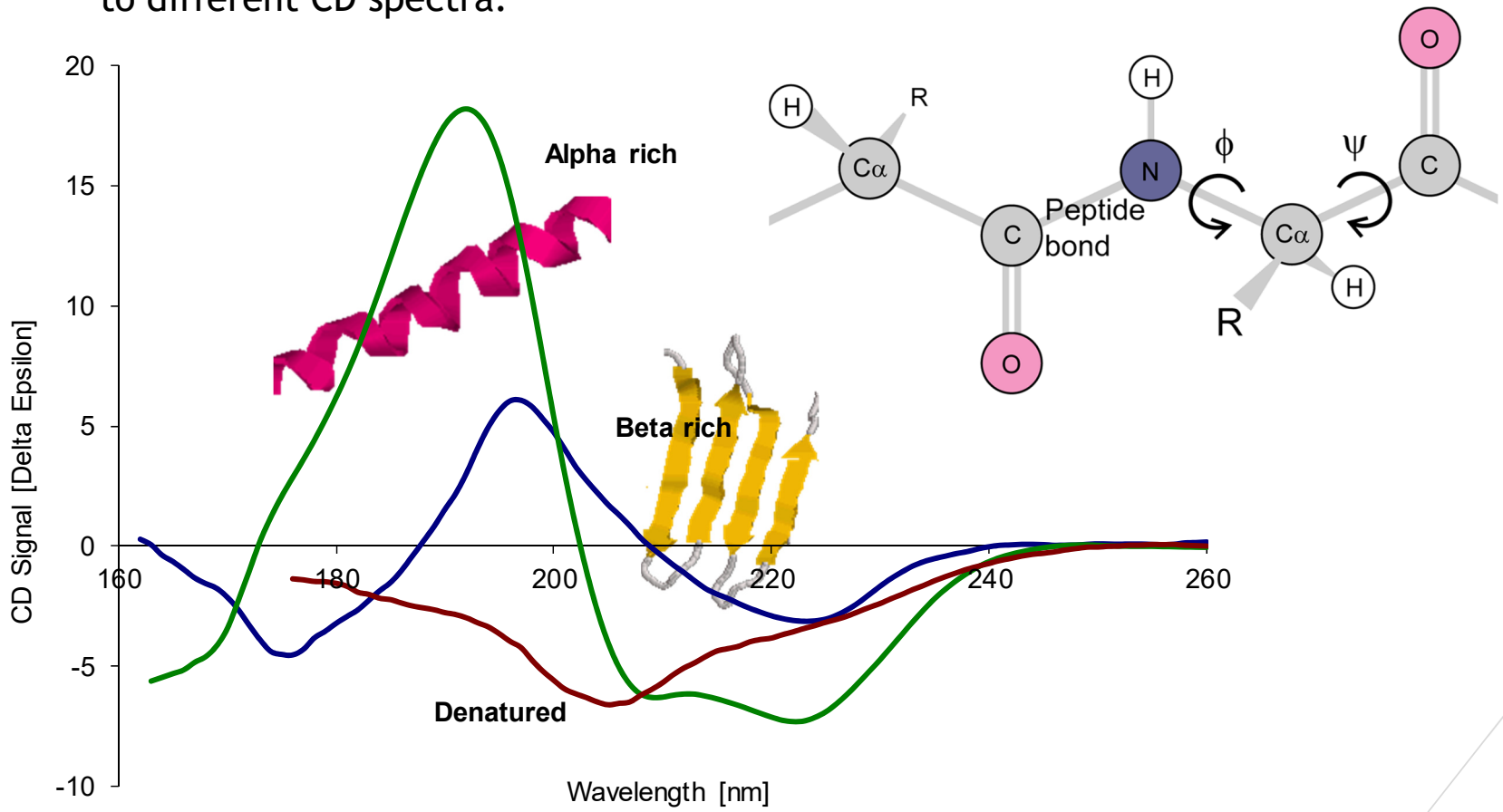
ESC1:
Circular Dichroism:
best practice and data analysis

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

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

Information in a protein CD spectrum

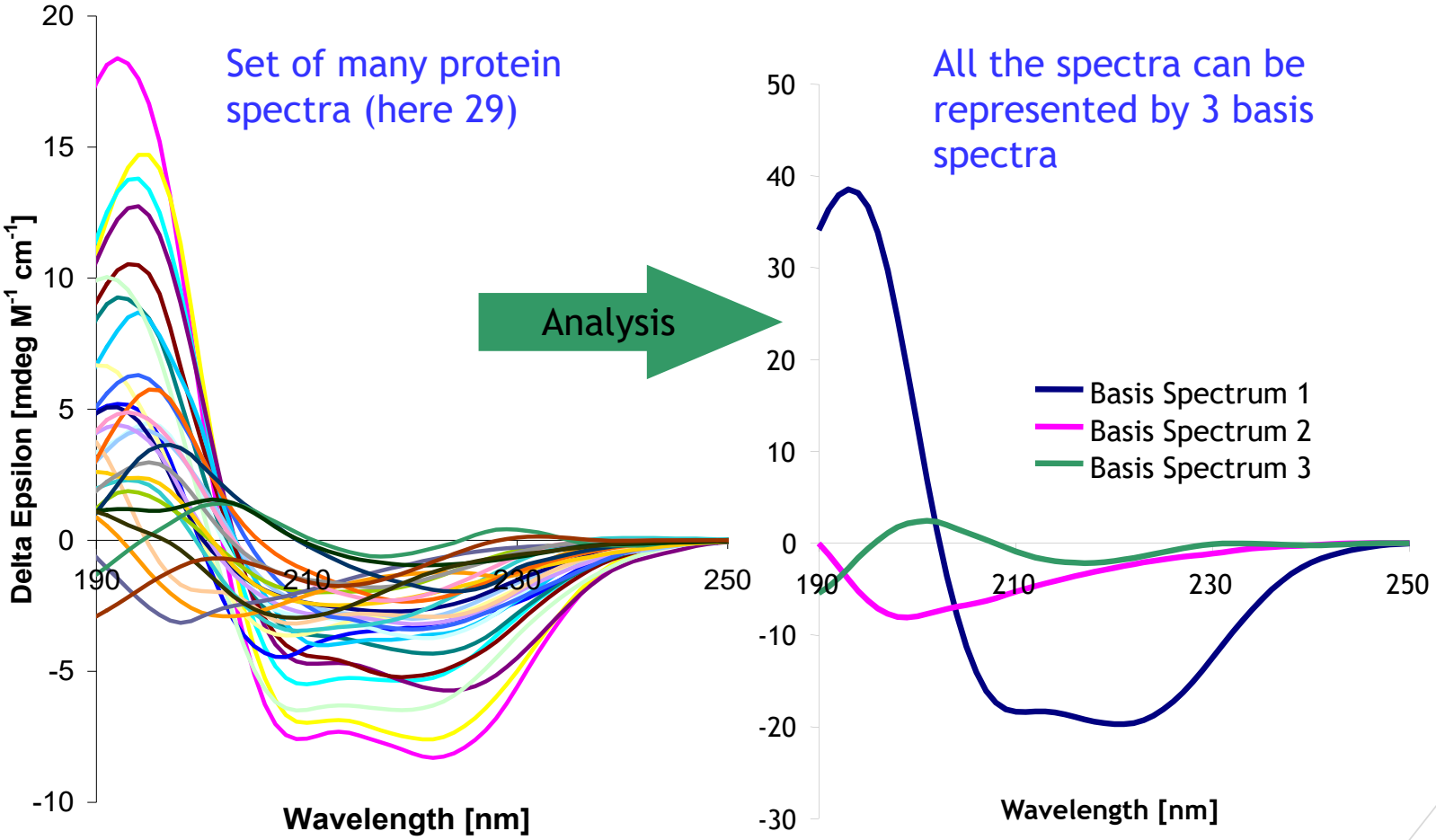
Different secondary structure in proteins gives rise to different CD spectra.



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

MOSBRI
Molecular-Scale Biophysics
Research Infrastructure

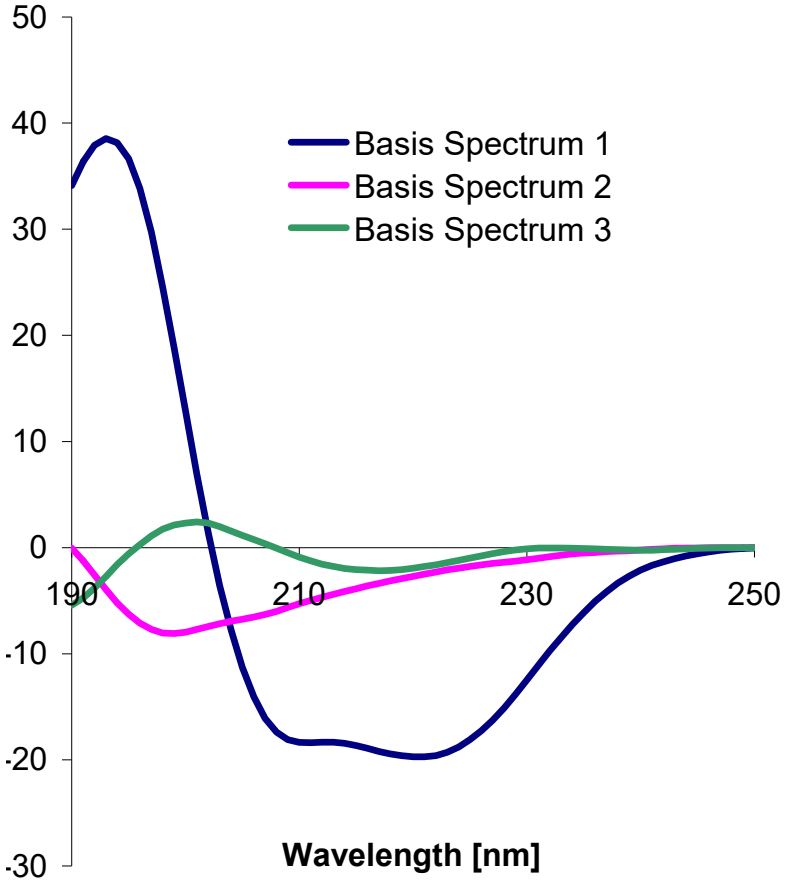
Information in a protein CD spectrum

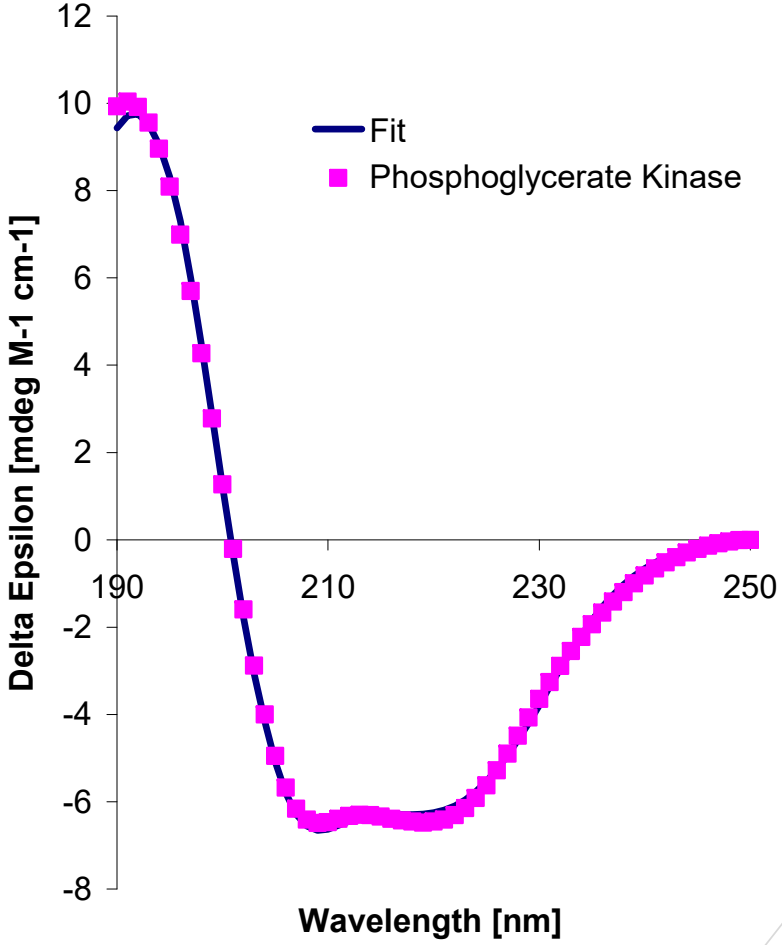


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004806

MOSBRI Molecular-Scale Biophysics Research Infrastructure

All spectra may be reconstructed from the three basis spectra





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

MOSBRI
Molecular-Scale Biophysics Research Infrastructure

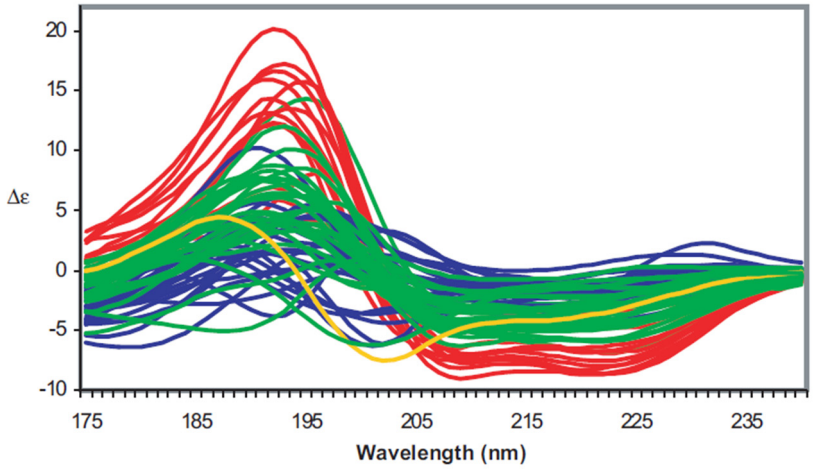
CD spectrum analysis

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	……..
α-helix	0.458	0.311	0.285	……..
β-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!

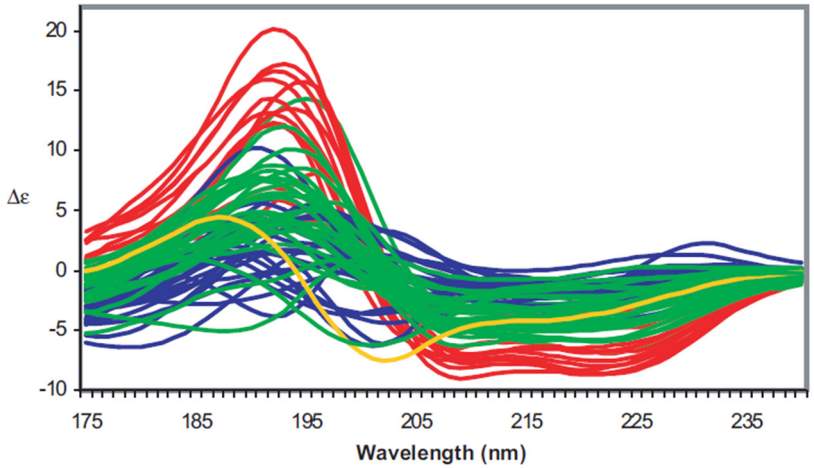
SP175 reference dataset: J.G. Lees *et al*. Bioinformatics. **22** (2006) 1955–1962

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

71 CD spectra of proteins



Write them as Matrices

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

m x 71 matrix

where m is number of data points

All the protein secondary structures (from crystallography)

	Aldolase	Alkaline phosphatase	Alpha amylase	.....
α-helix	0.458	0.311	0.285	.....
β-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!

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

l x 71 matrix

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

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

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

m x 71 matrix
where m is number of data points

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

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}*

Short cut

- The Singular Value Decomposition (SVD) theorem:
  U, S and V exists: $A = USV^T$ where $V^TV = UU^T = I$, 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: $\mathbf{V^T}$

X exists via Singular Value Decomposition (SVD)

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

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

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

$F = X\,A$ *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 α+B+T+O is in general not 100%
- Each component may even be < 0

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

$$f = \begin{bmatrix} \alpha \\ \beta \\ T \\ 0 \end{bmatrix} \quad \text{from} \quad A_{ref} = \begin{bmatrix} \text{CD of Aldolase} & \text{CD of Alkaline...} & \text{CD of Alpha...} & \cdots\cdots & \cdots\cdots & \text{CD of Ubiquitin} \end{bmatrix} \quad C = \begin{bmatrix} \text{CD of protein} \end{bmatrix}$$

$$F_{ref} = \begin{matrix} \text{Aldolase} & \text{Alkaline...} & & \text{Ubiquitin} \end{matrix} \begin{bmatrix} 0.458 & 0.311 & \cdots\cdots & 0.250 \\ 0.138 & 0.183 & & 0.316 \\ 0.103 & 0.133 & & 0.145 \\ 0.301 & 0.374 & & 0.289 \end{bmatrix}$$

Several methods exist which use SVD and finds the structure $f = \begin{bmatrix} \alpha \\ \beta \\ T \\ 0 \end{bmatrix}$ 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_{ref}.
- CONTIN/LL: Fit the spectrum C to a combination of spectra in A_{ref} resembling C the most

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

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

<u>As an example, Selcon3 does:</u>

- Sort A_{ref} such that the spectrum which resembles C the most are to the left
- Guess a solution f_{guess} (from the spectrum resembling C the most)
- Add C to A_{sort} and f_{guess} to F_{ref}: A = [C, A_{sort}] and F=[f_{guess}, F_{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_{guess}

Repeat the until self-consistency

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

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

Selcon3

1) Sort and guess $A_{sort} = [\text{CD Resemble C most}, \text{CD Resemble C less}, \ldots\ldots, \text{CD Resemble C least}]$, $F_{sort} = [\text{f Resemble C most}, \text{f Resemble C less}, \ldots\ldots, \text{f Resemble C least}]$, $f_{guess} = [\alpha_g, \beta_g, T_g, O_g] = [\text{f Resemble C most}]$

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

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

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

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

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

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

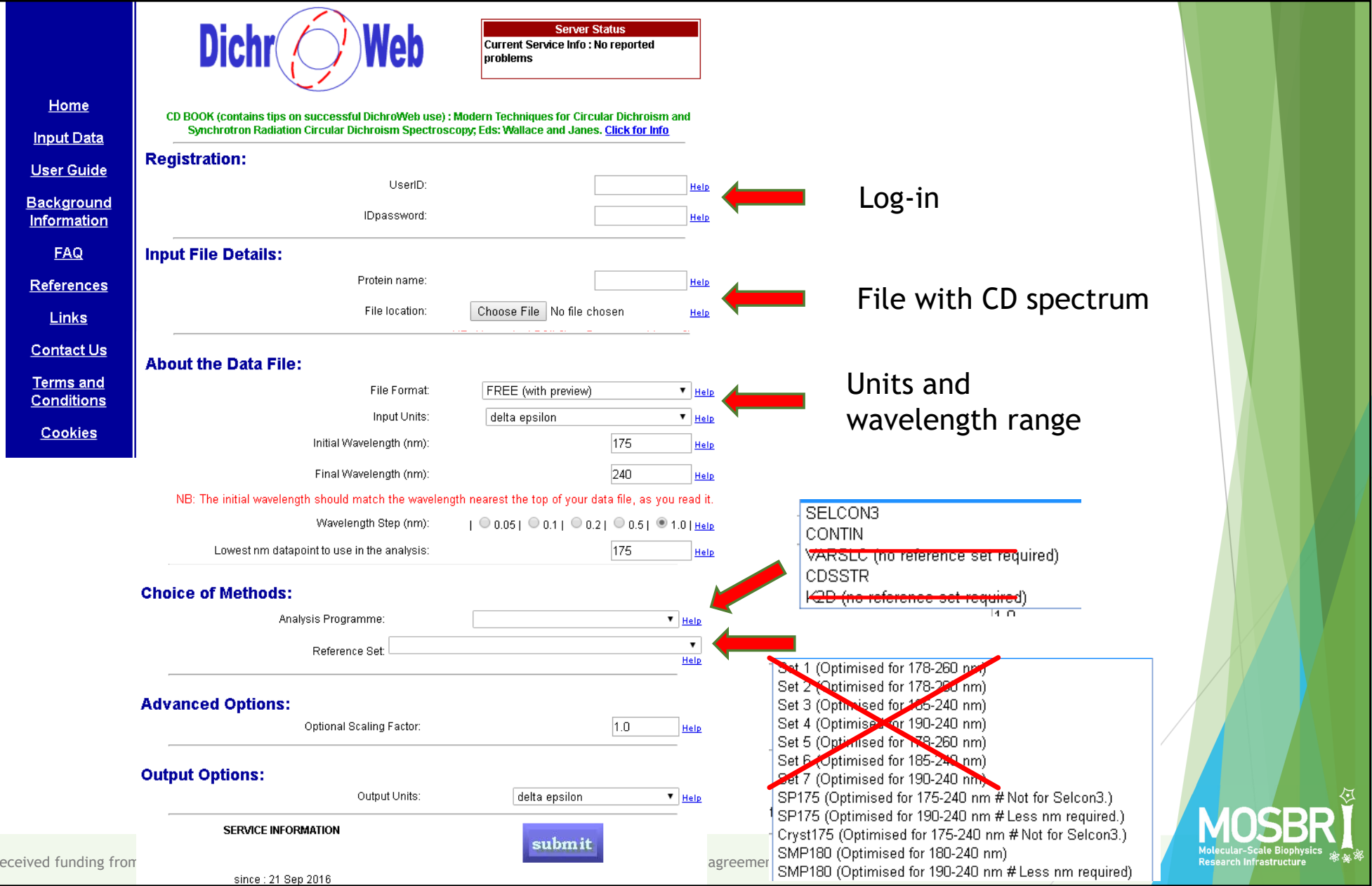
MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure


Home
Input Data
User Guide
Background Information
FAQ
References
Links
Contact Us
Terms and Conditions
Cookies
DichroWeb
Server Status
Current Service Info : No reported problems
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:
IDpassword:
Help
Input File Details:
Protein name:
File location:
Choose File
No file chosen
About the Data File:
File Format:
FREE (with preview)
Input Units:
delta epsilon
Initial Wavelength (nm):
175
Final Wavelength (nm):
240
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
Lowest nm datapoint to use in the analysis:
Choice of Methods:
Analysis Programme:
Reference Set:
Advanced Options:
Optional Scaling Factor:
1.0
Output Options:
Output Units:
delta epsilon
SERVICE INFORMATION
since : 21 Sep 2016
submit
Log-in
File with CD spectrum
Units and wavelength range
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)
This project has received funding from
agreemer
MOSBRI
Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis

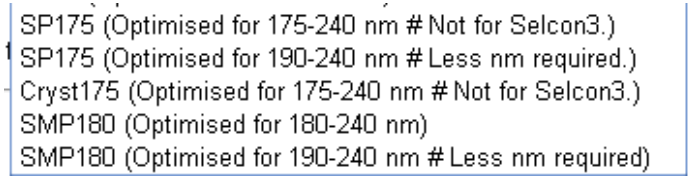
Reference CD data:

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

SP175: Soluble proteins

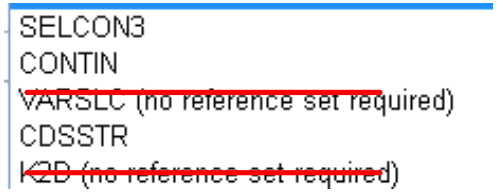
SMP180: Membrane proteins

Data down to 175 and 180 nm, respectively.



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 α-helix, β-sheet etc.

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis: Example

Human osteopontin (hOPN)

hOPN

Solutions from the CDSSTR method

Solutions using reference database: SP175.

Use of the reference set requries the citation of:
Lees, J.G., Miles, A.J., Wien, F., and Wallace, B.A. (2006), Bioinformatics, 22, 1955-1962.

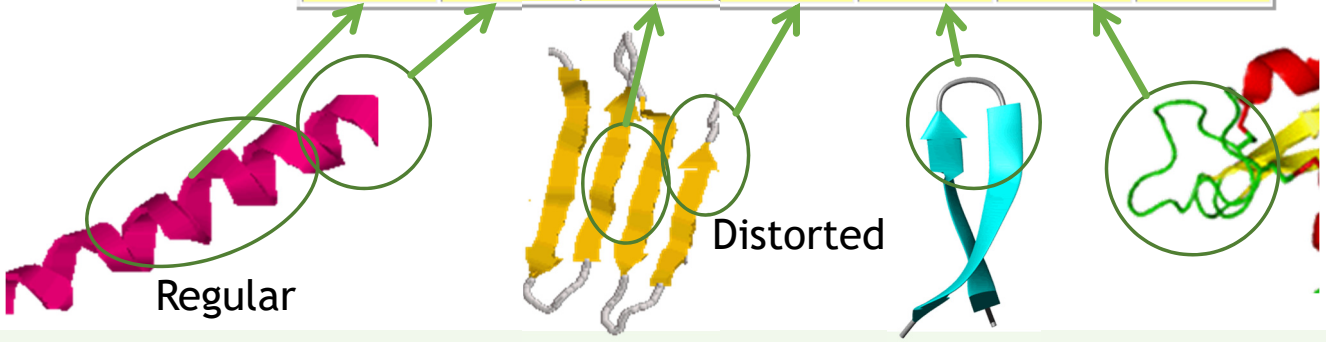
NRMSD:0.038

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



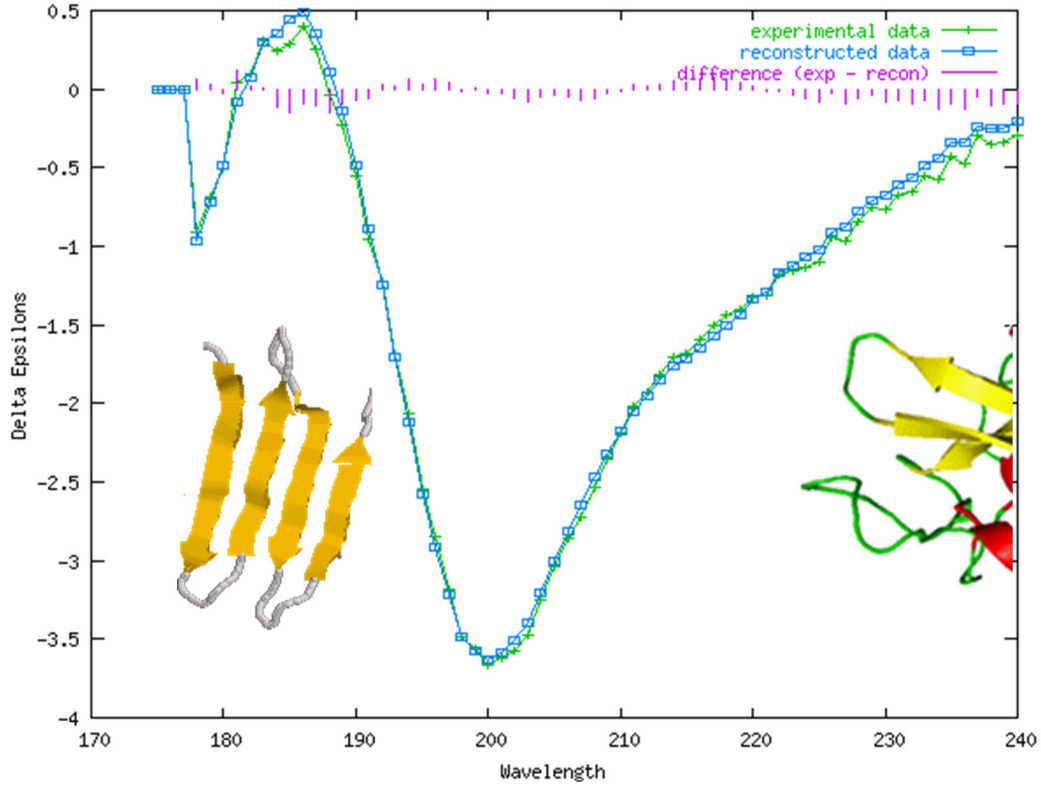
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004806

MOSBRI
Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis: Example

Human osteopontin (hOPN)

Dichroweb also displays a calculated CD spectrum for comparison.



Typical spectrum for a highly unordered protein.

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis: Example

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

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

Eva Kläning [a,b], Brian Christensen [a], Goran Bajic [a], Søren V. Hoffmann [c], Nykola C. Jones [c], Morten M. Callesen [a], Gregers R. Andersen [a], Esben S. Sørensen [a,d], Thomas Vorup-Jensen [b,d,e,*]

[a] Dept. of Molecular Biology and Genetics Aarhus University, Aarhus, Denmark
[b] Dept. of Biomedicine, Denmark
[c] Institute for Storage Ring Facilities Aarhus (ISA), Dept. of Physics and Astronomy & Center for Storage Ring Facilities Aarhus, Denmark
[d] Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Aarhus Denmark
[e] 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 α-helical[a], disordered α-helix[b], regular β-strand[c], disordered β-strand[d], turns[e] and unordered conformation[f] in OPN, dOPN, HCM and β-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 α-helix[a]	Distorted α-helix[b]	Regular β-strand[c]	Distorted β-strand[d]	Turns[e]	Unordered[f]	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

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

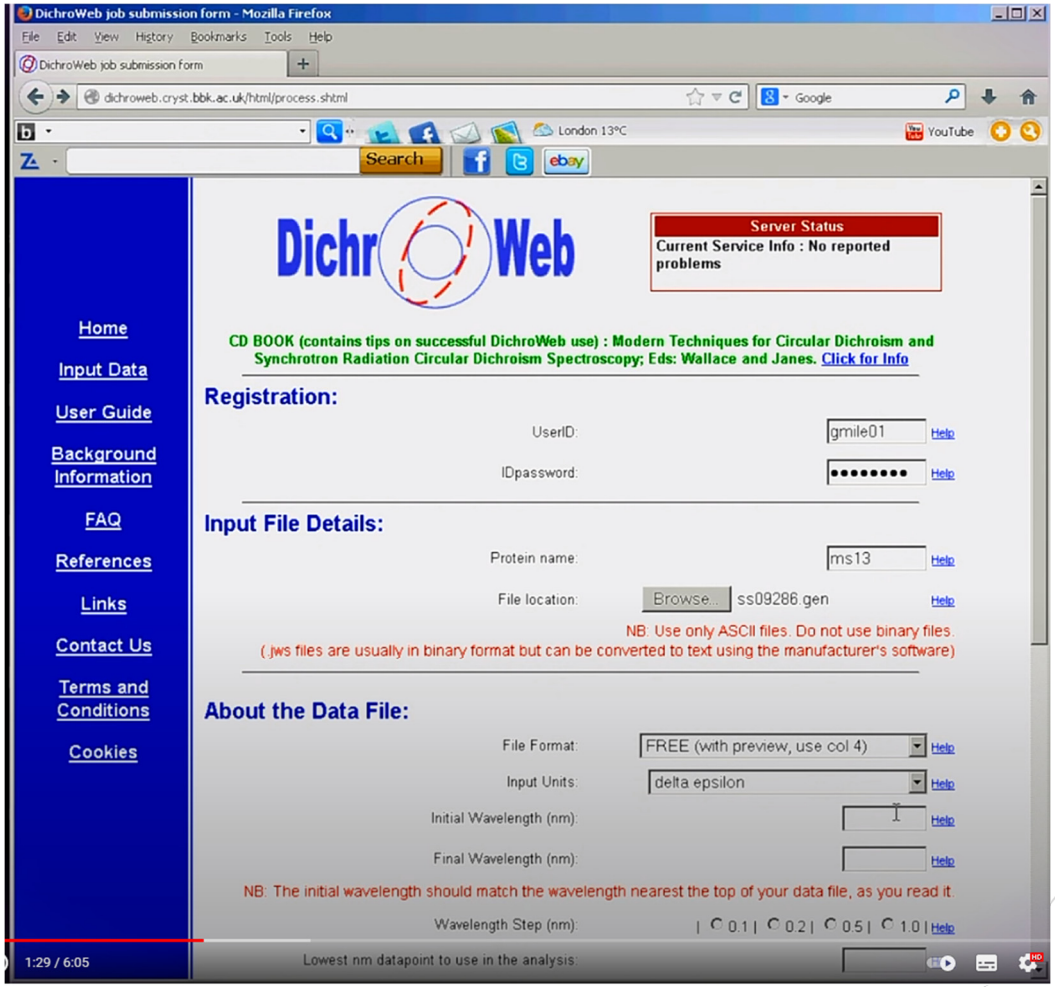
MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis: DichroWeb

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

ThePcddb
55 abonnenter

youtube.com/user/ThePcddb/videos



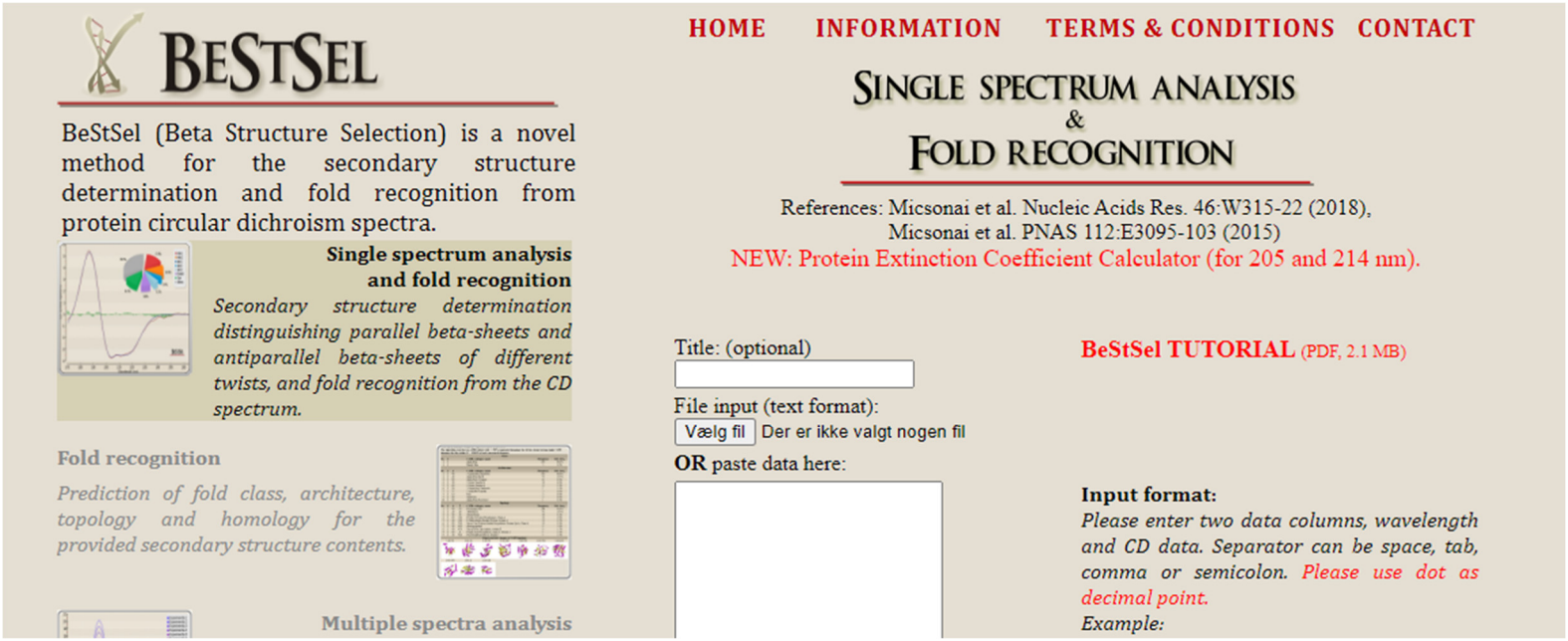
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004806

MOSBRI Molecular-Scale Biophysics Research Infrastructure

CD spectrum analysis: BestSel

Another good option for secondary structure analysis is BestSel

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



BestSel has more emphasis on different beta-sheet structures

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

MOSBRI Molecular-Scale Biophysics Research Infrastructure

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 an 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

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

MOSBRI
Molecular-Scale Biophysics Research Infrastructure