# TUTORIAL

latest update: 13-04-2025

Circular Dichroism (CD) spectroscopy is a widely used technique for the study of protein structure. Numerous algorithms have been developed for the estimation of the secondary structure composition from CD spectra. These methods often fail to provide acceptable results on α/β mixed or β-structure-rich proteins. The problem arises from the spectral diversity of β-structures. In Micsonai *et al.*, (2015) *Proc. Natl. Acad. Sci. USA* 112, E3095-E3103, we have shown that the parallel/antiparallel orientation and the twisting of β-sheets account for the observed spectral diversity. We developed the Beta Structure Selection (BeStSel) method for the secondary structure estimation that takes the twist of β-structures into account. This method can reliably distinguish parallel and antiparallel β-sheets and provides an improved secondary structure estimation for a broad range of proteins. Moreover, the secondary structure components applied by the method are characteristic to the protein fold and thus the fold can be predicted to the level of topology in the CATH classification (Orengo *et al.,* (1997) *Structure* 5(8):1093-1108.) from a single CD spectrum.

In publications using BeStSel method for secondary structure analysis, please kindly cite Micsonai *et al.*, (2015) *Proc. Natl. Acad. Sci. USA* 112, E3095-E3103 and Micsonai *et al*., (2018) *Nucleic Acids Res.* 46, W315–W322, or Micsonai et al. *Nucleic Acids Res.* 50, W90-98 (2022).

Here, we provide a brief introduction for the use of the BeStSel web server https://bestsel.elte.hu. The server is under continuous development. Although we make all efforts for its perfect functioning, we do not take the responsibility for any prediction error or software problems. We highly appreciate any questions or suggestions on the use of the server or reports on bugs found. Please, feel free to send us a message through the homepage (Contact page) or by email to kardos@elte.hu or micsonai@ttk.elte.hu.

For all details on the BeStsel method, beyond this tutorial, please, see the *Information* provided on the web server pages and refer to the original publications of Micsonai et al.

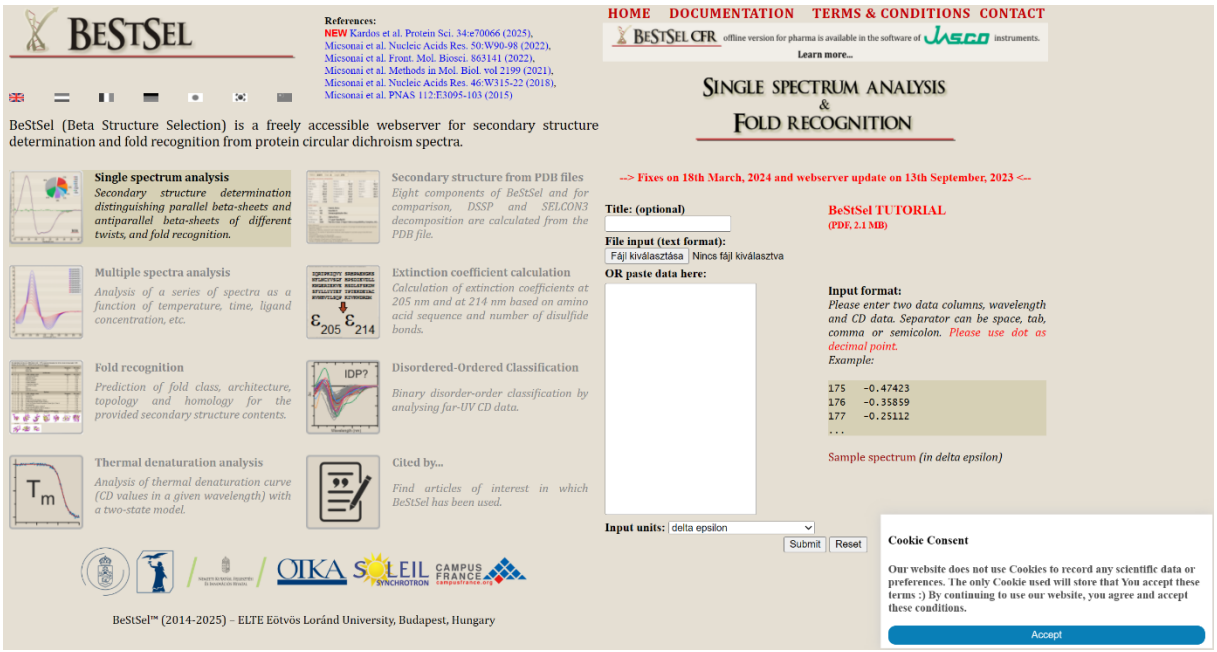
# Table of contents

Introduction ..... 3
Single spectrum analysis ..... 4
Input units ..... 4
Results format ..... 7
Data in text ..... 8
Wavelength range, scale factor, best factor ..... 10
Fold recognition ..... 12
Note on data smoothing and NRMSD ..... 12
Multiple spectra analysis ..... 13
Fold recognition ..... 17
Thermal denaturation analysis ..... 22
Data upload ..... 22
Fitting parameters ..... 25
Secondary structure from PDB files ..... 26
Extinction coefficient calculator ..... 29
Disordered-ordered classification ..... 30
Cited by… ..... 32
Documentation ..... 33

# Introduction

First, one of the 8 modules of the server can be chosen, listed on the left side of the starting page: *Single spectrum analysis, Multiple spectra analysis, Fold recognition, Thermal denaturation analysis, Secondary structure from PDB files, Extinction coefficient calculator, Disordered-ordered classification* and *Cited by...*

A language selector located in the top left corner is currently under development. In the future, the user will be able to select from English, Hungarian, French, German, Japanese, Korean or Chinese. Currently the web page is only available in English.



# Single spectrum analysis

In *Single spectrum analysis*, a single CD spectrum can be analyzed for the secondary structure composition and the protein fold can be predicted.

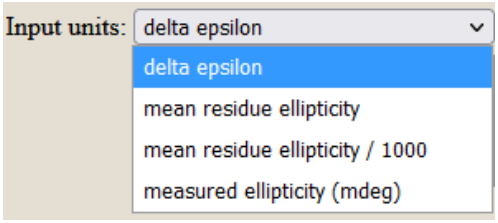
Data can be uploaded from a text file or can be copied into the window in two data columns, separator can be space, tab, comma or semicolon. Please use dot as decimal point. In case of browsed data file in text format, the system automatically recognizes the header and the data columns.

Input units

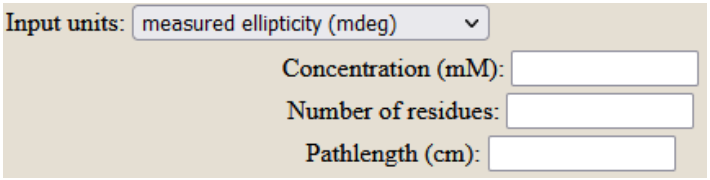
You can choose the appropriate Input units from the pop-up menu:

**Delta epsilon** ($M^{-1}$ $cm^{-1}$)

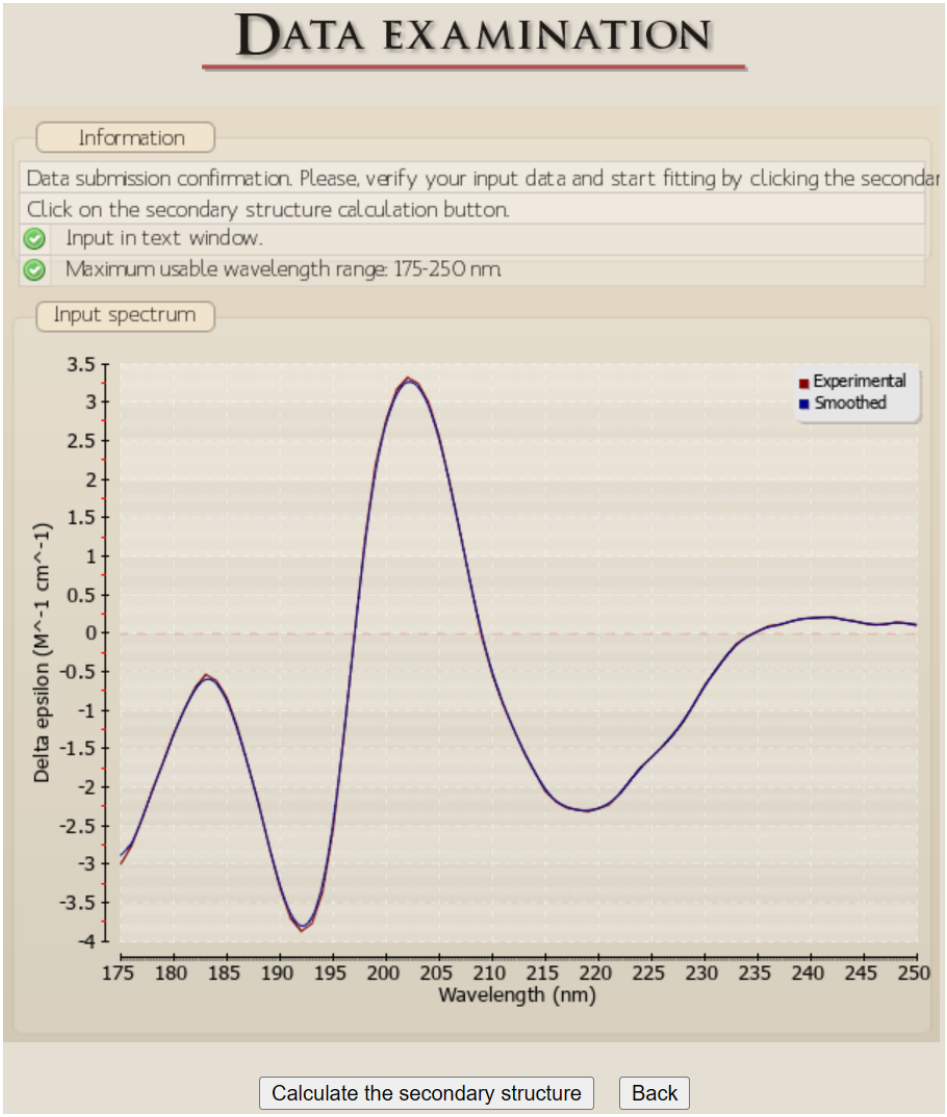
**Mean residue molar ellipticity** (deg $cm^2$ $dmol^{-1}$). $([\theta]_{MRW} = \theta/(10 \times c_r \times l)$, where $c_r$ is the molar concentration per residue, l is the pathlength.




**Measured ellipticity** data can be directly uploaded. In that case, the protein concentration in µM, the number of residues per protein molecule and the pathlength in cm should be provided by the user.



A *Data examination* window will appear to check if the data was uploaded properly. Data is converted to delta epsilon. Raw and smoothed data (data average on 2 nm window) are shown. The system makes an automatic data examination and gives warning note in case of unexpected CD amplitudes (calculation is still possible).

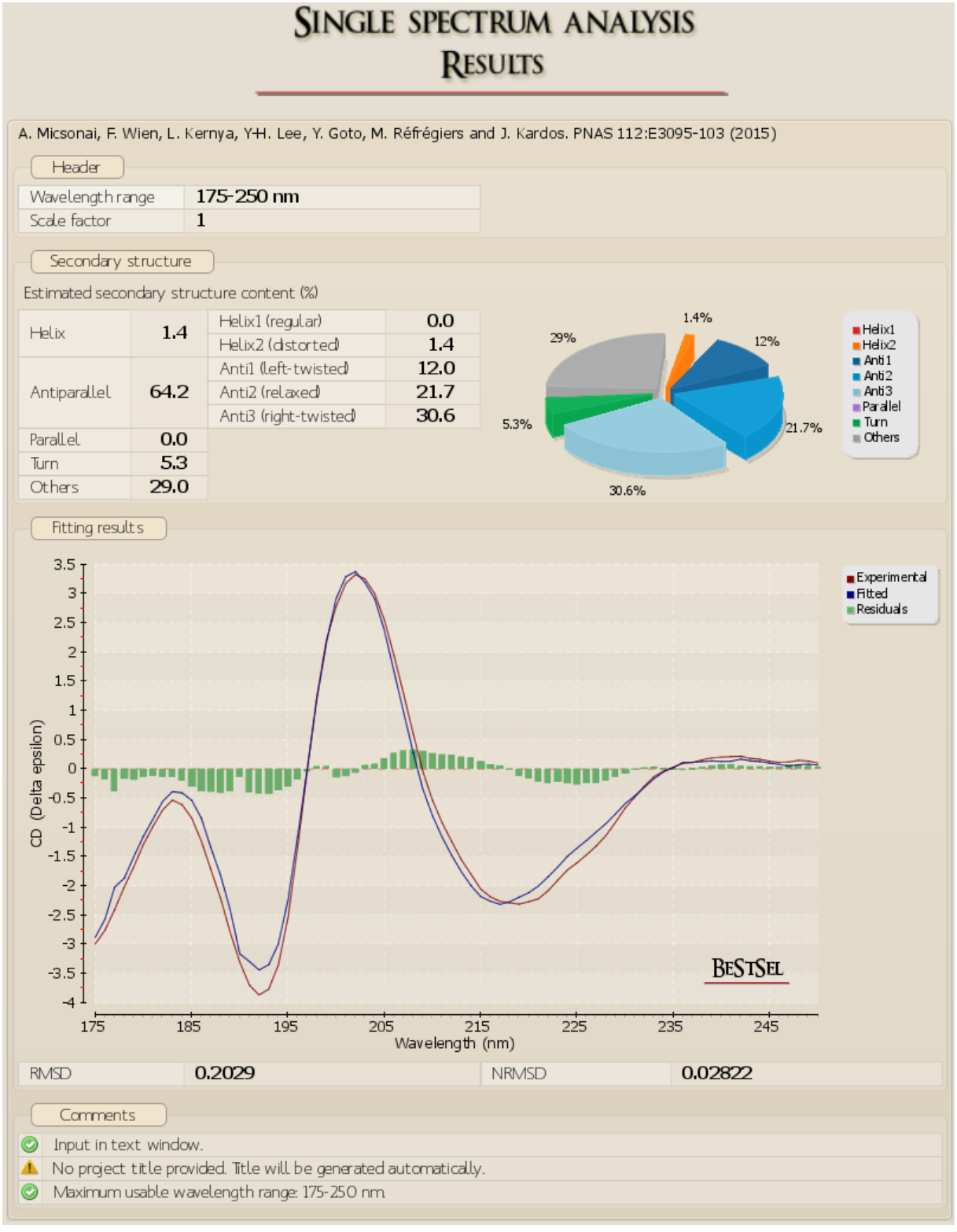



Please check carefully the wavelength range and amplitude of the CD data.

Secondary structure calculation can be initiated by clicking on "*Calculate the secondary structure*" bottom.

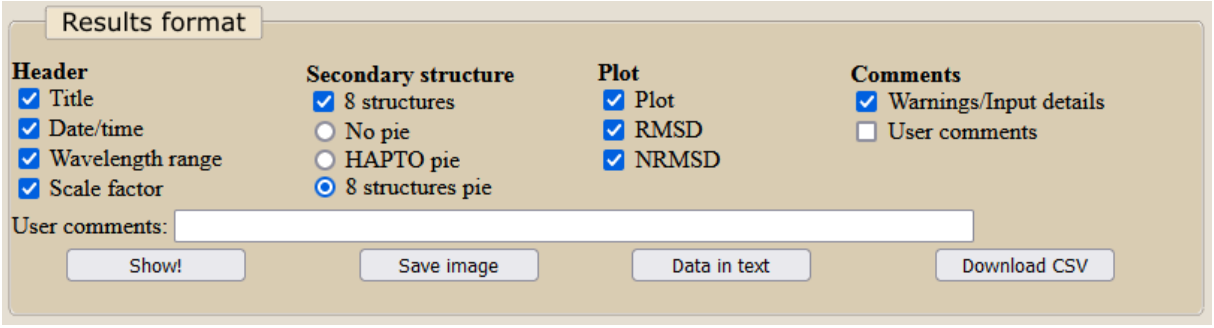
In the results window, the results will appear in a graphical image with all the useful information provided (including wavelength range and user-provided information). At first, data is analyzed in the possible widest wavelength range of the uploaded data. In general, a wider wavelength range contains more spectral information and a better accuracy is anticipated. However, we strongly suggest to choose an appropriate wavelength range where the PMT voltage was below the instrument limit (e.g., 600 volts) upon the measurement.

BeStSel calculates the contents for 8 secondary structure elements. For information on the secondary structure components and the fitting processes, please, refer to the original publications of Micsonai *et al*. (open access publications, link are provided on the front page of the webserver).



Results format

Below the results (please, roll down if it is not on the screen), the output format can be changed for the convenience of the user.




By choosing "Show!" the *Results* page can be reformatted. "*Save image"* will open the results in a separate browser window and can be saved as an image.

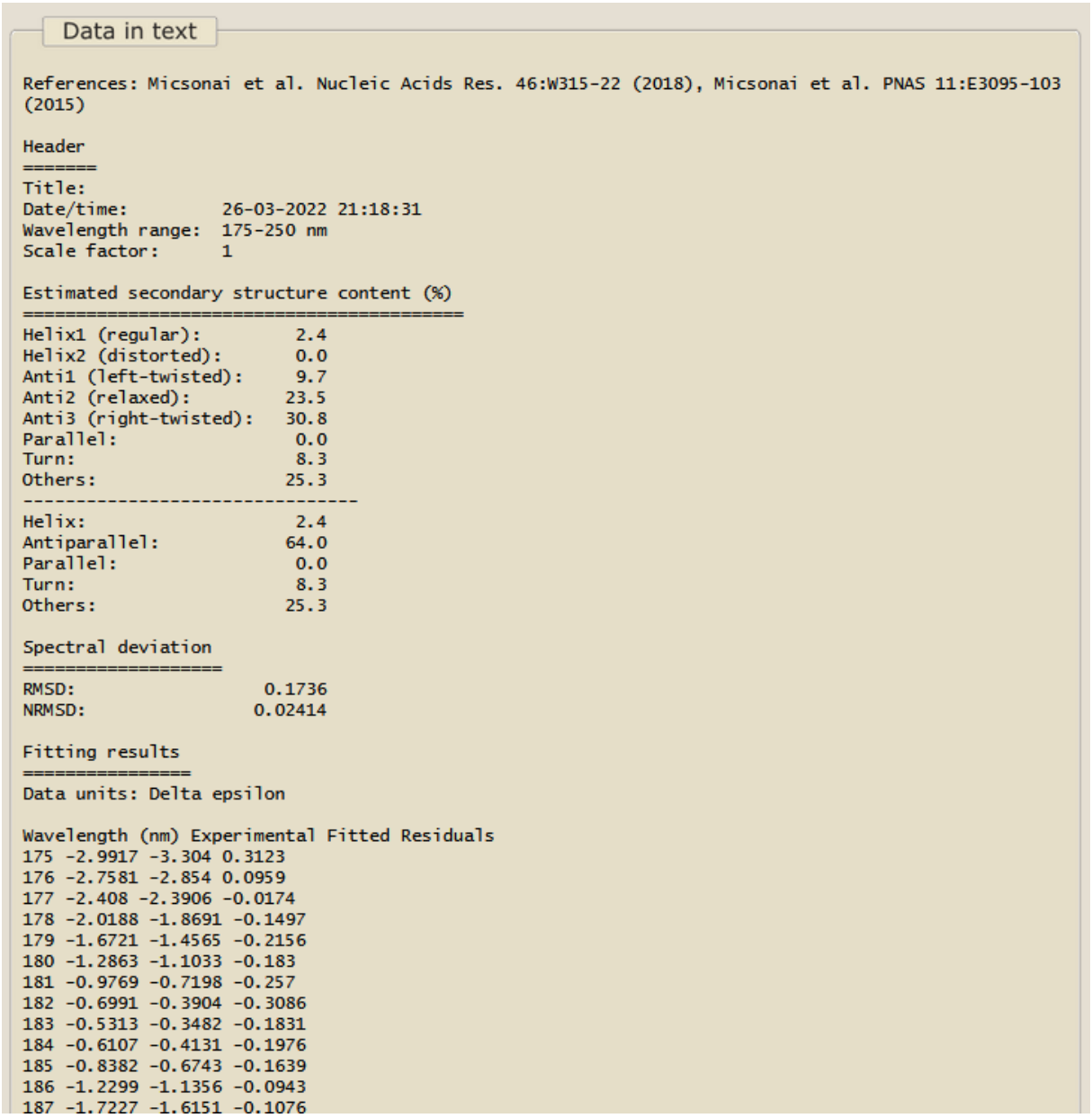
RMSD: root mean square deviation. $\sqrt{\frac{1}{w}\sum_{i=1}^{w}\left(CD_{exp,i}-CD_{fit,i}\right)^2}$

NRMSD: normalized root mean square deviation.
$\frac{1}{\max(CD_{exp})-\min(CD_{exp})}\sqrt{\frac{1}{w}\sum_{i=1}^{w}\left(CD_{exp,i}-CD_{fit,i}\right)^2}$

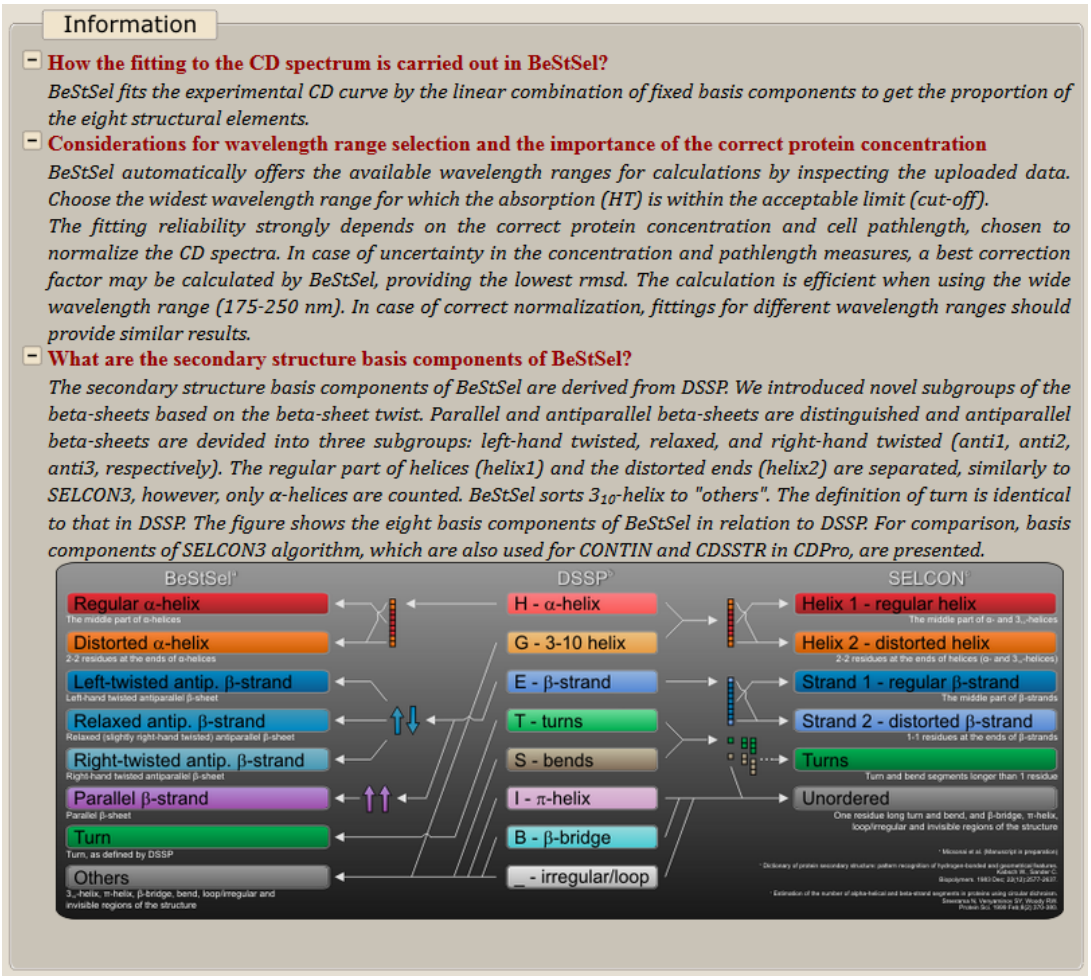
From 13th September, 2023, NRMSD is calculated between the smoothed experimental spectrum (2 nm window, smoothing by data averaging) and the fitted spectrum.

Data in text

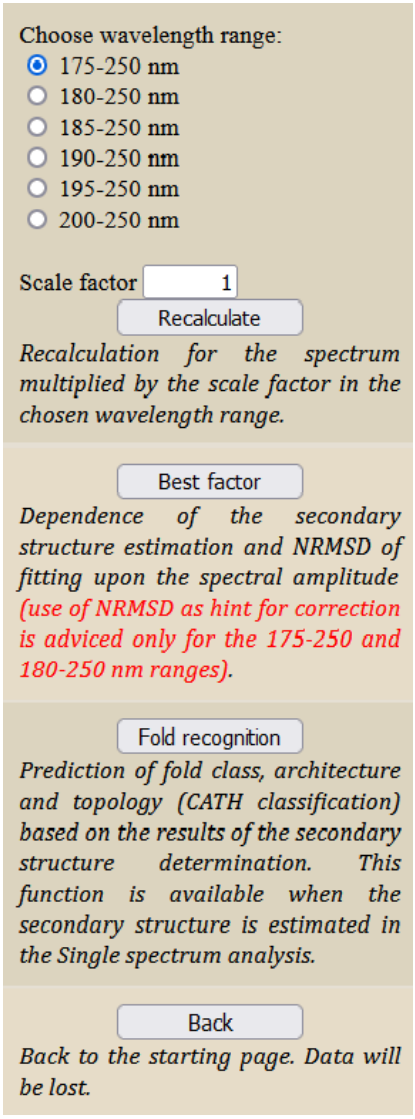
For further data processing by the users, result can be shown in text format with the predicted results at the top and the experimental, fitted, and the residual data in columns below. By copying, the data can be transferred to any data processing software to make your own plots, etc.



At the bottom of the *Results* page, brief information on the BeStSel fitting and some advices to consider are provided.



Wavelength range, scale factor, best factor




On the left side of the *Results* page, the wavelength range can be chosen and the analysis can be recalculated. A scale factor can be chosen for the recalculation as well. The CD amplitude is multiplied with this factor.

The "*Best factor"* function carries out a series of analyses by changing the current scaling factor automatically in the range of 0.5-2.

The factor related to the lowest NRMSD is highlighted (see next page). The dependence of the individual secondary structure components on the CD amplitude is plotted. This can be informative in the case of uncertainties in the protein concentration or pathlength. For CD data in a wide wavelength range (down to at least 180 nm), a change in the factor from 1 to the lowest fitting NRMSD is an indicator of incorrect concentration or pathlength values.

Please note that the automatic scaling calculation of Best factor shows the dependence of the secondary structure estimation and NRMSD on the amplitude of your spectrum. The factor with the lowest NRMSD should not be taken as correction for your normalized spectrum when used in the 190-250 or 200-250 nm range. The correct concentration determination is essential for accurate analysis. When 175-250 or 180-250 range is used and the Best factor is significantly different from 1.0, it indicates possible normalization problems, and the factor can be taken as suggestion.

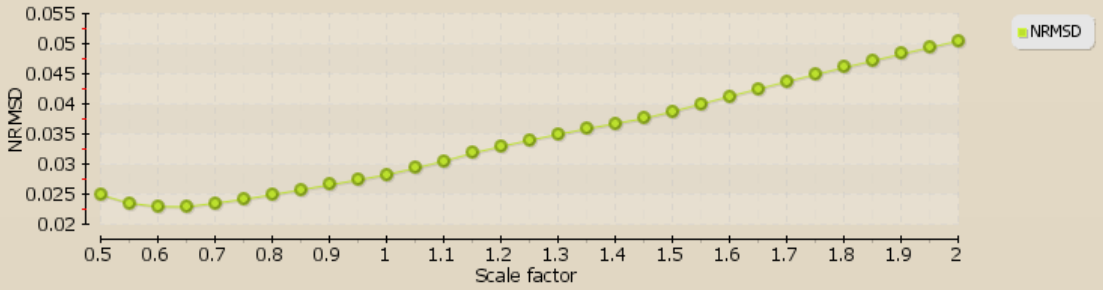
The "*Best factor*" results can be saved as an image or in text format by giving the format of the results at the bottom of the page.

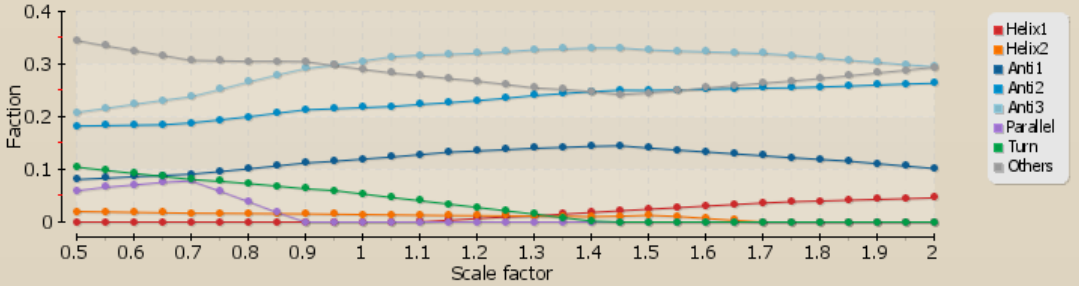
# BEST FACTOR RESULTS

A. Micsonai, F. Wien, L. Kernya, Y-H. Lee, Y. Goto, M. Réfrégiers and J. Kardos. PNAS 112:E3095-103 (2015)

Header

| | |
|---|---|
| Wavelength range | 175-250 nm |
| Scale factor | 1 |




| factor | Helix1 % | Helix2 % | Anti1 % | Anti2 % | Anti3 % | Parallel % | Turn % | Others % | NRMSD |
|---|---|---|---|---|---|---|---|---|---|
| 0.50 | 0.00 | 1.99 | 8.07 | 18.25 | 20.83 | 5.95 | 10.44 | 34.47 | 0.0249 |
| 0.55 | 0.00 | 1.91 | 8.33 | 18.30 | 21.57 | 6.59 | 9.82 | 33.47 | 0.0235 |
| 0.60 | 0.00 | 1.83 | 8.59 | 18.39 | 22.35 | 7.05 | 9.24 | 32.56 | 0.0229 |
| 0.65 | 0.00 | 1.76 | 8.85 | 18.47 | 23.09 | 7.53 | 8.66 | 31.65 | 0.0229 |
| 0.70 | 0.00 | 1.68 | 9.09 | 18.83 | 23.80 | 7.81 | 8.09 | 30.71 | 0.0235 |
| 0.75 | 0.00 | 1.65 | 9.58 | 19.39 | 25.16 | 5.88 | 7.73 | 30.62 | 0.0241 |
| 0.80 | 0.00 | 1.61 | 10.09 | 19.99 | 26.58 | 3.89 | 7.31 | 30.53 | 0.0249 |
| 0.85 | 0.00 | 1.59 | 10.65 | 20.66 | 27.90 | 1.83 | 6.83 | 30.54 | 0.0257 |
| 0.90 | 0.00 | 1.55 | 11.21 | 21.32 | 29.15 | 0.00 | 6.31 | 30.47 | 0.0265 |

Fold recognition

Protein fold can be predicted from the results of the CD spectrum analysis. For information on this method please see the “Fold recognition” module in this tutorial, or the *Information* on the main BeStSel page.

Note on data smoothing and NRMSD

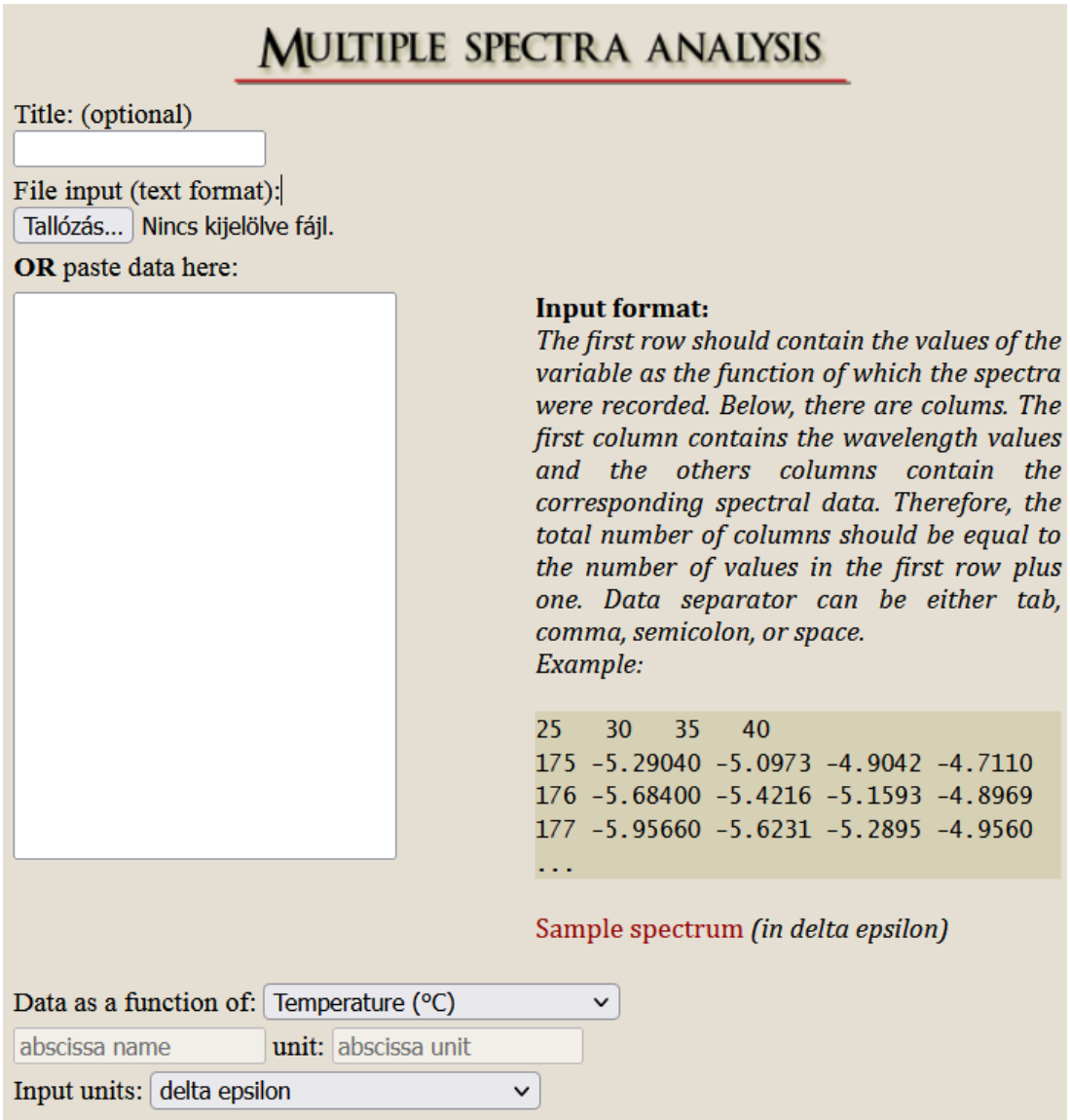
The updated version of BeStSel applies data smoothing with 2 nm window prior to secondary structure analysis. This smoothing window has no significant effect on the analyses of noiseless, smooth spectra. However, in the case of noisy spectra, which often occur in the lower wavelength range, this smoothing increases the performance of the method. In case our users do smoothing themselves, we do not recommend a stronger smoothing (i.e. using a larger window) because that will significantly distort the spectrum and affect the secondary structure estimation.

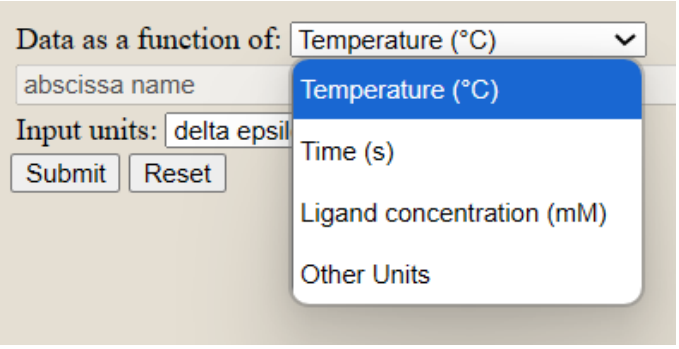
In general, we suggest to record the CD spectra with 0.1 nm data pitch in as high quality as possible by increasing the recording time of data points in step scan mode or increasing the number of accumulated scans in continuous scanning mode. Please, note that BeStSel will provide you the fitting with 1 nm data step.

NRMSD: In the case of noisy spectra, there will be always larger RMSD and NRMSD values, calculated between the input and the fitted spectra, than in the case of a noiseless spectrum, even if BeStSel has a similar performance. In such a situation, the RMSD and NRMSD values will be largely proportional to the noise and will not reflect correctly the reliability of the BeStSel fitting. To show how good the fitting is, instead of the input spectrum, we use the smoothed spectrum vs. the fitted one for NRMSD calculation. For RMSD, we kept the calculation between the original experimental spectrum and the fitted one, thus, RMSD is highly sensitive for noise level.

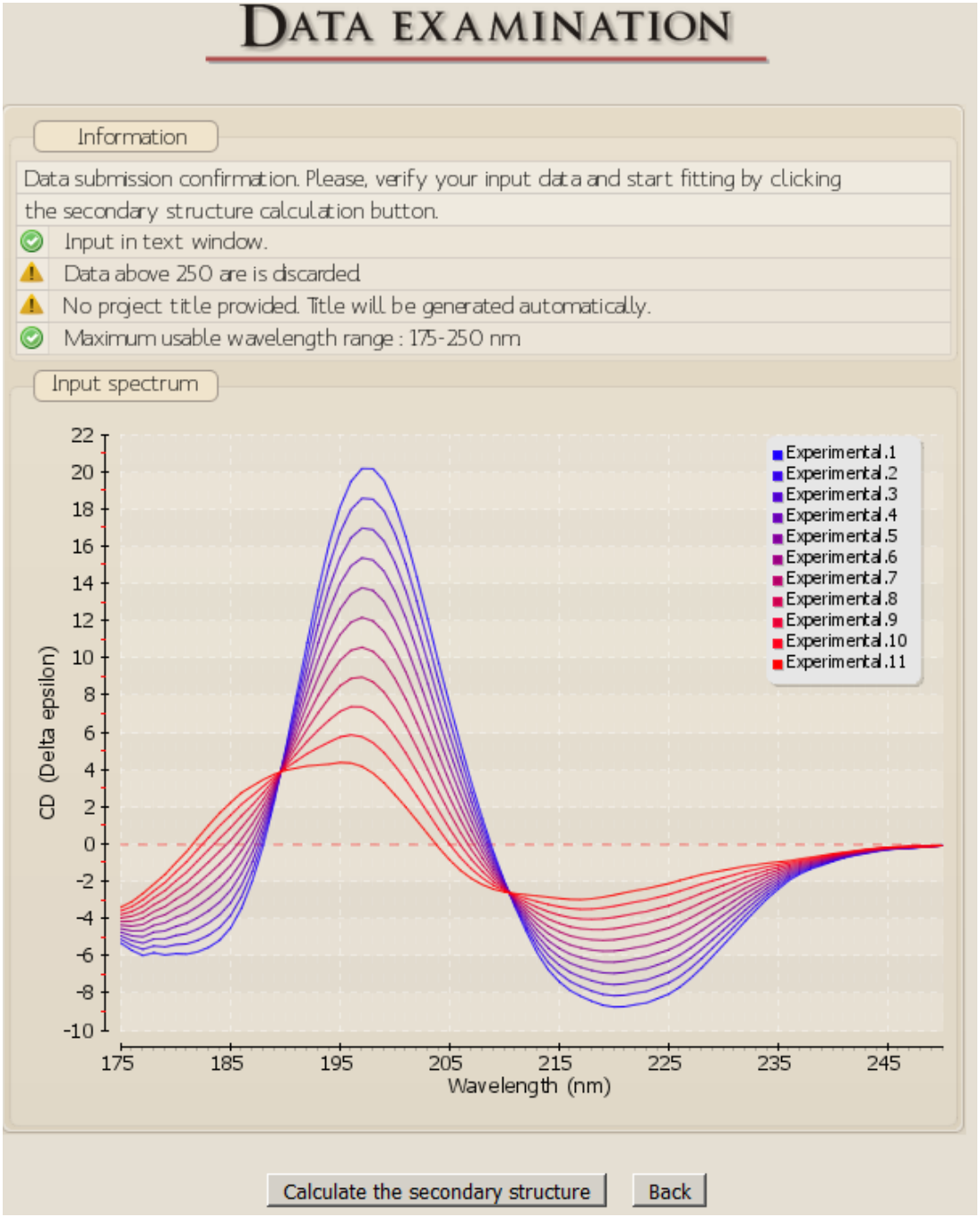
# Multiple spectra analysis

A series of spectra can be uploaded from a file or copied into the window from a worksheet. The first row should contain the values of the parameter that was varied in the measurements (e. g. temperature values, if the spectra were recorded at different temperatures). Below, there are columns. The first column contains the wavelength values and the others columns contain the corresponding spectral data. Therefore, the total number of columns should be equal to the number of values in the first row plus one. Data separator can be either tab, comma, semicolon or space.

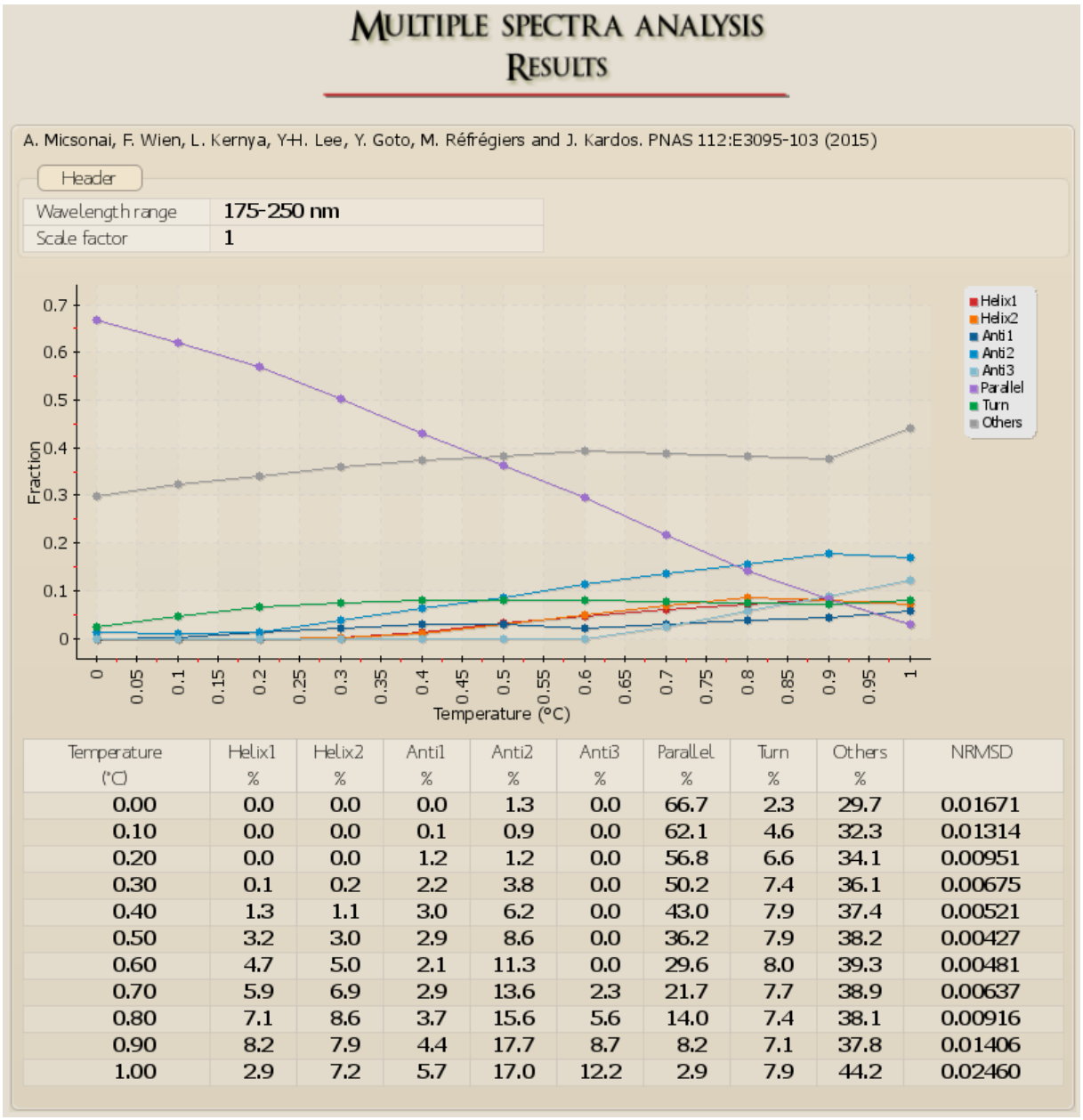





First, a data examination page comes up to check if the upload was correct. Then, all the spectra are evaluated at the same time and shown as a function of the chosen parameter.

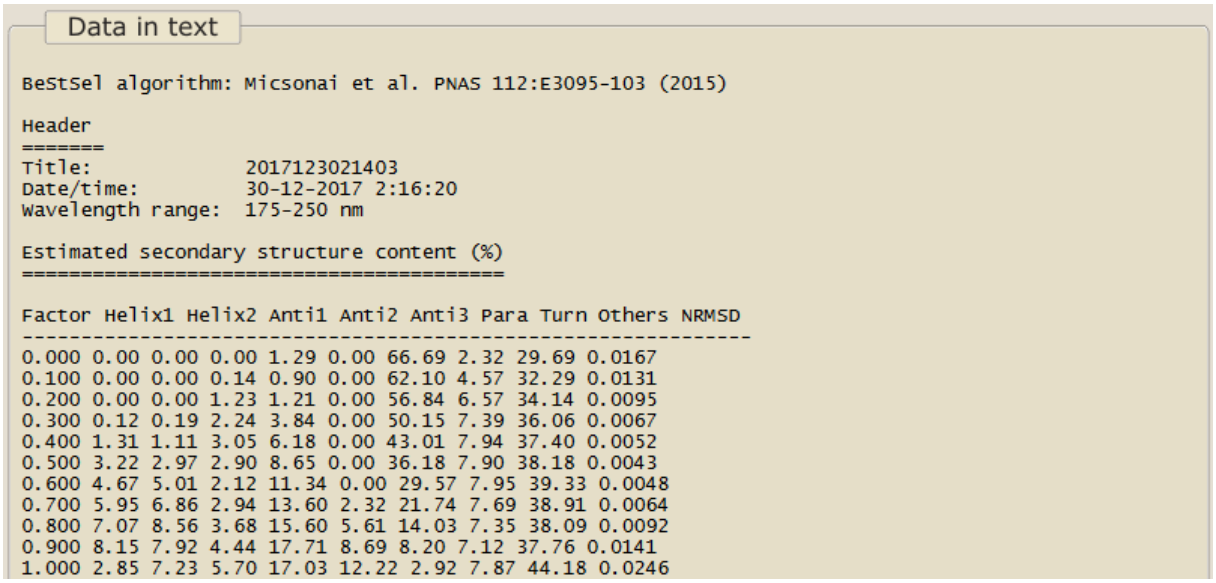


After clicking on the "*Calculate the secondary structure*" button, the result window will appear.




| Temperature (°C) | Helix1 % | Helix2 % | Anti1 % | Anti2 % | Anti3 % | Parallel % | Turn % | Others % | NRMSD |
|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.0 | 0.0 | 0.0 | 1.3 | 0.0 | 66.7 | 2.3 | 29.7 | 0.01671 |
| 0.10 | 0.0 | 0.0 | 0.1 | 0.9 | 0.0 | 62.1 | 4.6 | 32.3 | 0.01314 |
| 0.20 | 0.0 | 0.0 | 1.2 | 1.2 | 0.0 | 56.8 | 6.6 | 34.1 | 0.00951 |
| 0.30 | 0.1 | 0.2 | 2.2 | 3.8 | 0.0 | 50.2 | 7.4 | 36.1 | 0.00675 |
| 0.40 | 1.3 | 1.1 | 3.0 | 6.2 | 0.0 | 43.0 | 7.9 | 37.4 | 0.00521 |
| 0.50 | 3.2 | 3.0 | 2.9 | 8.6 | 0.0 | 36.2 | 7.9 | 38.2 | 0.00427 |
| 0.60 | 4.7 | 5.0 | 2.1 | 11.3 | 0.0 | 29.6 | 8.0 | 39.3 | 0.00481 |
| 0.70 | 5.9 | 6.9 | 2.9 | 13.6 | 2.3 | 21.7 | 7.7 | 38.9 | 0.00637 |
| 0.80 | 7.1 | 8.6 | 3.7 | 15.6 | 5.6 | 14.0 | 7.4 | 38.1 | 0.00916 |
| 0.90 | 8.2 | 7.9 | 4.4 | 17.7 | 8.7 | 8.2 | 7.1 | 37.8 | 0.01406 |
| 1.00 | 2.9 | 7.2 | 5.7 | 17.0 | 12.2 | 2.9 | 7.9 | 44.2 | 0.02460 |

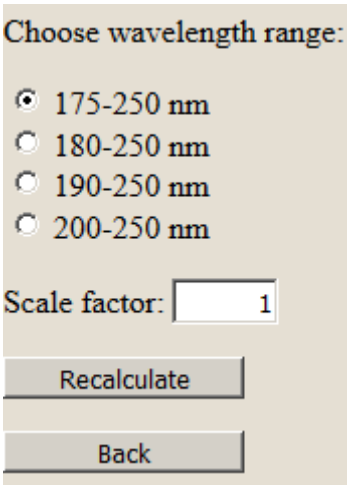
At the bottom, the image can be chosen to be saved and is opened in a separate window. Also, results in text format can be chosen for further data processing by the user.



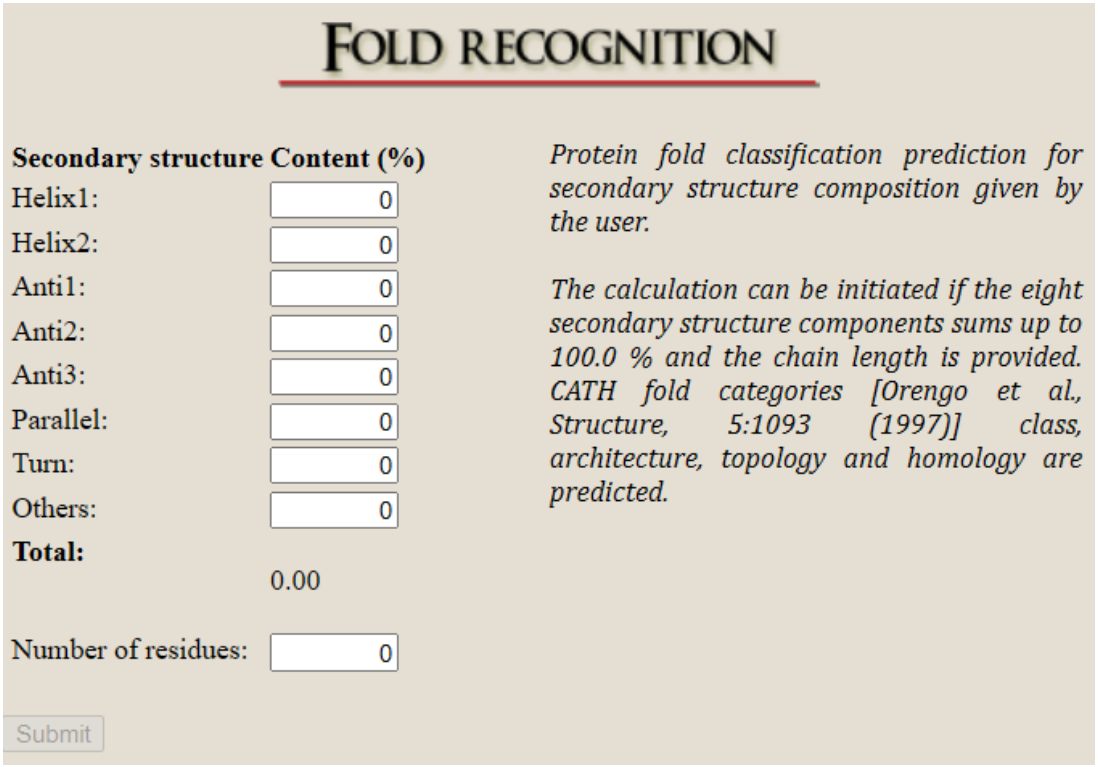
Data in text

```
BeStSel algorithm: Micsonai et al. PNAS 112:E3095-103 (2015)

Header
=======
Title:              2017123021403
Date/time:          30-12-2017 2:16:20
Wavelength range:   175-250 nm

Estimated secondary structure content (%)
=========================================

Factor Helix1 Helix2 Anti1 Anti2 Anti3 Para Turn Others NRMSD
-------------------------------------------------------------
0.000 0.00 0.00 0.00 1.29 0.00 66.69 2.32 29.69 0.0167
0.100 0.00 0.00 0.14 0.90 0.00 62.10 4.57 32.29 0.0131
0.200 0.00 0.00 1.23 1.21 0.00 56.84 6.57 34.14 0.0095
0.300 0.12 0.19 2.24 3.84 0.00 50.15 7.39 36.06 0.0067
0.400 1.31 1.11 3.05 6.18 0.00 43.01 7.94 37.40 0.0052
0.500 3.22 2.97 2.90 8.65 0.00 36.18 7.90 38.18 0.0043
0.600 4.67 5.01 2.12 11.34 0.00 29.57 7.95 39.33 0.0048
0.700 5.95 6.86 2.94 13.60 2.32 21.74 7.69 38.91 0.0064
0.800 7.07 8.56 3.68 15.60 5.61 14.03 7.35 38.09 0.0092
0.900 8.15 7.92 4.44 17.71 8.69 8.20 7.12 37.76 0.0141
1.000 2.85 7.23 5.70 17.03 12.22 2.92 7.87 44.18 0.0246
```

On the left side, the wavelength range can be changed or a scaling factor can be set and the data can be re-analyzed.



# Fold recognition




The „Fold recognition" module of the server is used to predict the fold of a protein structure from the secondary structure contents. The calculation can be initiated if the eight secondary structure components sum up to 100.0 % and the chain length is provided. These data may come from previous BeStSel analysis of a CD spectrum (see „Single spectrum analysis" module) or from the analysis of a PDB structure (see „Secondary structure from PDB files" module).

4 different analyses are provided: (1) a search for similar structures on the entire PDB, (2) a fold search on the closest structures on a non-redundant single domain PDB subset, (3) a search on single domains with secondary structure composition within the expected error of the CD secondary structure analysis and (4) a weighted K-nearest neighbors search method.

## FOLD RECOGNITION RESULTS

**Closest structures in the entire PDB**
**The twenty closest structures based on the euclidean distance in the eight dimensional space of BeStSel.**

| Secondary structure content (%) | BeStSel | 1VTZ | 3RQV | 2JWP | 3KGZ | 3S4G |
|---|---|---|---|---|---|---|
| Helix1 | 1.4 | 3.4 | 3.4 | 1.1 | 3.5 | 2.6 |
| Helix2 | 7.8 | 6.1 | 6.1 | 6.9 | 5.1 | 6.1 |
| Anti1 | 5.6 | 4.0 | 3.9 | 6.4 | 7.8 | 3.0 |
| Anti2 | 23.4 | 24.6 | 24.5 | 21.0 | 23.4 | 25.9 |
| Anti3 | 12.0 | 12.8 | 13.0 | 14.6 | 11.7 | 12.9 |
| Para | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Turn | 9.9 | 8.8 | 8.8 | 9.8 | 9.3 | 9.7 |
| Others | 39.9 | 40.4 | 40.3 | 40.2 | 39.1 | 39.8 |
| **Molecular description** | | | | | | |
| Num. of res. | | 5880 | 5880 | 174 | 312 | 196 |
| Missing res. | | 360 | 360 | 0 | 22 | 12 |
| Num. of chains | | 30 | 30 | 1 | 2 | 1 |
| **CATH data** | | | | | | |
| Link to CATH | | 1VTZ | 3RQV | 2JWP | 3KGZ | 3S4G |
| CATH Domains | | 6.1.1* (10)<br>2.60.120(20) | 2.60.120(30) | 6.1.1* (1) | NA | NA |

*: 6.1.1. means Unclassified CATH Domain.
Numbers in brackets indicate the number of a given CATH domain in the structure.

**Closest single CATH domains in a PDB subset filtered for <= 90% sequence homology.**
**The ten closest structures based on the euclidean distance in the eight dimensional space of BeStSel.**

| No. | PDB | Chain | C | A | T | Class | Architecture | Topology | Length |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1VTZ | 2 | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 196 |
| 2 | 1J58 | A | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 384 |
| 3 | 1JHJ | A | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 171 |
| 4 | 1G6E | A | 2 | 60 | 20 | Mainly Beta | Sandwich | Gamma-B Crystallin; domain 1 | 87 |
| 5 | 2PHL | A | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 397 |
| 6 | 2W2A | A | 2 | 40 | 128 | Mainly Beta | Beta Barrel | Lipocalin | 194 |
| 7 | 1LED | A | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 243 |
| 8 | 1ZKO | A | 2 | 40 | 50 | Mainly Beta | Beta Barrel | OB fold (Dihydrolipoamide Acetyltransferase, E2P) | 135 |
| 9 | 1QWR | A | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 319 |
| 10 | 1B35 | B | 2 | 60 | 120 | Mainly Beta | Sandwich | Jelly Rolls | 255 |

**Structural domains within the expected error of BeStSel**

**The algorithm searches on a PDB subset with <= 90% sequential homology for all the chains having single CATH domains that lie within 1.5 × RMSD distance in each structural element from the estimated secondary structure.**

**Class**

| No. | C | CATH category name | Frequency | Rel. freq. |
|---|---|---|---|---|
| 1 | 2 | Mainly Beta | 26 | 96.3% |
| 2 | 3 | Alpha Beta | 1 | 3.7% |

**Architecture**

| No. | C | A | CATH category name | Frequency | Rel. freq. |
|---|---|---|---|---|---|
| 1 | 2 | 60 | Sandwich | 20 | 74.1% |
| 2 | 2 | 40 | Beta Barrel | 6 | 22.2% |
| 3 | 3 | 40 | 3-Layer(aba) Sandwich | 1 | 3.7% |

**Topology**

| No. | C | A | T | CATH category name | Frequency | Rel. freq. | Length* |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 60 | 120 | Jelly Rolls | 17 | 63.0% | 297 |
| 2 | 2 | 40 | 128 | Lipocalin | 5 | 18.5% | 184 |
| 3 | 2 | 60 | 20 | Gamma-B Crystallin; domain 1 | 2 | 7.4% | 87 |
| 4 | 2 | 60 | 40 | Immunoglobulin-like | 1 | 3.7% | 144 |
| 4 | 2 | 40 | 50 | OB fold (Dihydrolipoamide Acetyltransferase, E2P) | 1 | 3.7% | 135 |
| 4 | 3 | 40 | 1000 | Protein Transport Mog1p; Chain A | 1 | 3.7% | 177 |

*: Average chain length of the members of the topology in the box.

**Representative images of CATH topologies**

2.60.120 2.40.128 2.60.20 2.60.40 2.40.50 3.40.1000

For the weighted K- nearest neighbors method, the number of residues is required for the analysis.

**Weighted K-nearest neighbors search**

Chain length: 167 WKNN analysis

| Class | | | |
|---|---|---|---|
| **No.** | **CATH** | **Cathegory name** | **Score** |
| 1 | 2 | Mainly Beta | 17.34 |
| **Architecture** | | | |
| **No.** | **CATH** | **Cathegory name** | **Score** |
| 1 | 2.60 | Sandwich | 15.74 |
| 2 | 2.40 | Beta Barrel | 8.96 |
| 3 | 2.30 | Roll | 3.04 |
| 4 | 3.30 | 2-Layer Sandwich | 4.27 |
| 5 | 3.70 | Box | 1.49 |
| **Topology** | | | |
| **No.** | **CATH** | **Cathegory name** | **Score** |
| 1 | 2.60.120 | Jelly Rolls | 14.11 |
| 2 | 2.60.40 | Immunoglobulin-like | 3.23 |
| 3 | 2.40.128 | Lipocalin | 6.78 |
| 4 | 2.30.31 | Transcriptional Co-activator pc4; Chain A | 2.31 |
| 5 | 3.30.530 | Alpha-D-Glucose-1,6-Bisphosphate; Chain A, domain 4 | 3.60 |
| 6 | 3.70.10 | Proliferating Cell Nuclear Antigen | 1.49 |
| 7 | 2.40.160 | Porin | 2.22 |
| 8 | 2.60.200 | Tumour Suppressor Smad4 | 2.16 |
| 9 | 3.40.1000 | Protein Transport Mog1p; Chain A | 2.86 |
| 10 | 2.30.110 | Pnp Oxidase; Chain A | 1.44 |
| **Homology** | | | |
| **No.** | **CATH** | **Cathegory name** | **Score** |
| 1 | 2.60.120.430 | Galactose-binding lectin | 1.79 |
| 2 | 2.60.120.10 | Jelly Rolls | 12.31 |
| 3 | 2.60.40.840 | N/A | 1.69 |
| 4 | 2.60.120.20 | N/A | 2.45 |
| 5 | 2.40.128.20 | N/A | 6.08 |
| 6 | 2.60.40.1290 | N/A | 1.56 |
| 7 | 2.30.31.10 | Transcriptional Coactivator Pc4; Chain A | 2.31 |
| 8 | 3.30.530.20 | N/A | 2.24 |
| 9 | 2.60.120.260 | Galactose-binding domain-like | 2.96 |
| 10 | 2.60.40.740 | N/A | 2.23 |
| 11 | 3.70.10.10 | N/A | 1.49 |
| 12 | 2.40.160.90 | N/A | 2.22 |
| 13 | 2.60.200.40 | N/A | 2.16 |
| 14 | 3.40.1000.10 | Mog1/PsbP, alpha/beta/alpha sandwich | 2.86 |
| 15 | 2.30.110.20 | Hcp1-like | 1.44 |

The weighted K- nearest neighbors method predict the Class, Architecture, Topology, and Homology of the protein using the single domain subset of CATH 4.3 (see the number of domains and categories in the table below). In each layer (Class, Architecture, Topology, Homology) the predicted categories are ordered by their calculated WKNN scores excluding every structure that belongs to an already predicted category (lower numbered hits). The WKNN score is defined by the sum of the weighted distance of every structure (from the query point) among the K- nearest neighbors which belong to that particular category.

| Number of | CATH 4.3 |
|---|---|
| Domains | 61932 |
| Class | 5 |
| Architecture | 43 |
| Topology | 1467 |
| Homology | 6540 |

At the bottom of the Fold recognition results, information on the analysis methods is provided.

Information

**How are the closest structures found in the entire PDB?**

*The BeStSel algorithm characterizes the secondary struncture of proteins by using eight components. Every single protein structure can be represented by a point in this eight dimensional secondary structural space. The distance between two points (x and x') is defined by their euclidean distance:*

$$d = \sqrt{\sum_{i=1}^{8} (x_i - x_i')^2}$$

*where $x_i$ is the content of the $i^{th}$ secondary structure of protein x. The search for the closest structures is carried out on the entire PDB of 71430 structures. The closest structures are presented by their PDB ID, secondary structure content, sequence length, number of chains and CATH domains contained. Links are provided for the PDB and CATH databases. This function is especially useful in case of multidomain proteins.*

**How are the closest single CATH domains found in the filtered PDB subset?**

*In case of single domain proteins, a prediction for CATH domains can be done. One option is the search for the closest structures on a single domain structure set. The definition of the distance is presented in the topic "How are the closest structures found in the entire PDB?". The single domain PDB subset is a non-redundant collection of chains containing single CATH domains or homodomains filtered for <=90% sequence homology and resolution better than 3.0 Angstr�ms. This dataset contains 9218 domains covering 4 classes, 38 architectures, and 764 topologies.*

*The method of searching for the closest structures does not take into acount the possible error of the secondary structure estimation. It can be used even if the secondary structural space is rarely populated by structures around the estimated point.*

**How are the structural domains found within the expected error of BeStSel?**

*The algorithm searches on a PDB subset with <= 90% sequential homology for all the chains having single CATH domains that lie within 1.5 × RMSD distance in each structural element from the estimated secondary structure. In other words we look up the structures in a box centered to the BeStSel result. The size of the box is determined by the RMSD of BeStSel on SP175 reference set. The hits in the box are sorted out for classes, architectures and topologies. The resulting table shows the frequences and percentages of the different groups in the CATH categories in the order of frequency. In cases of architecture and topology the ten most populated groups are presented.*

*In the most dense regions of the secondary structural space hundreds of points can be found in the box. The advantage of the method is that the rarely populated folds located within the error of estimation that do not appear with the "closest structure" method will also be surveyed. In the case of some unique secondary structure composition no point or only a few points are located in the box. The closest structures method can be useful in such a situation.*

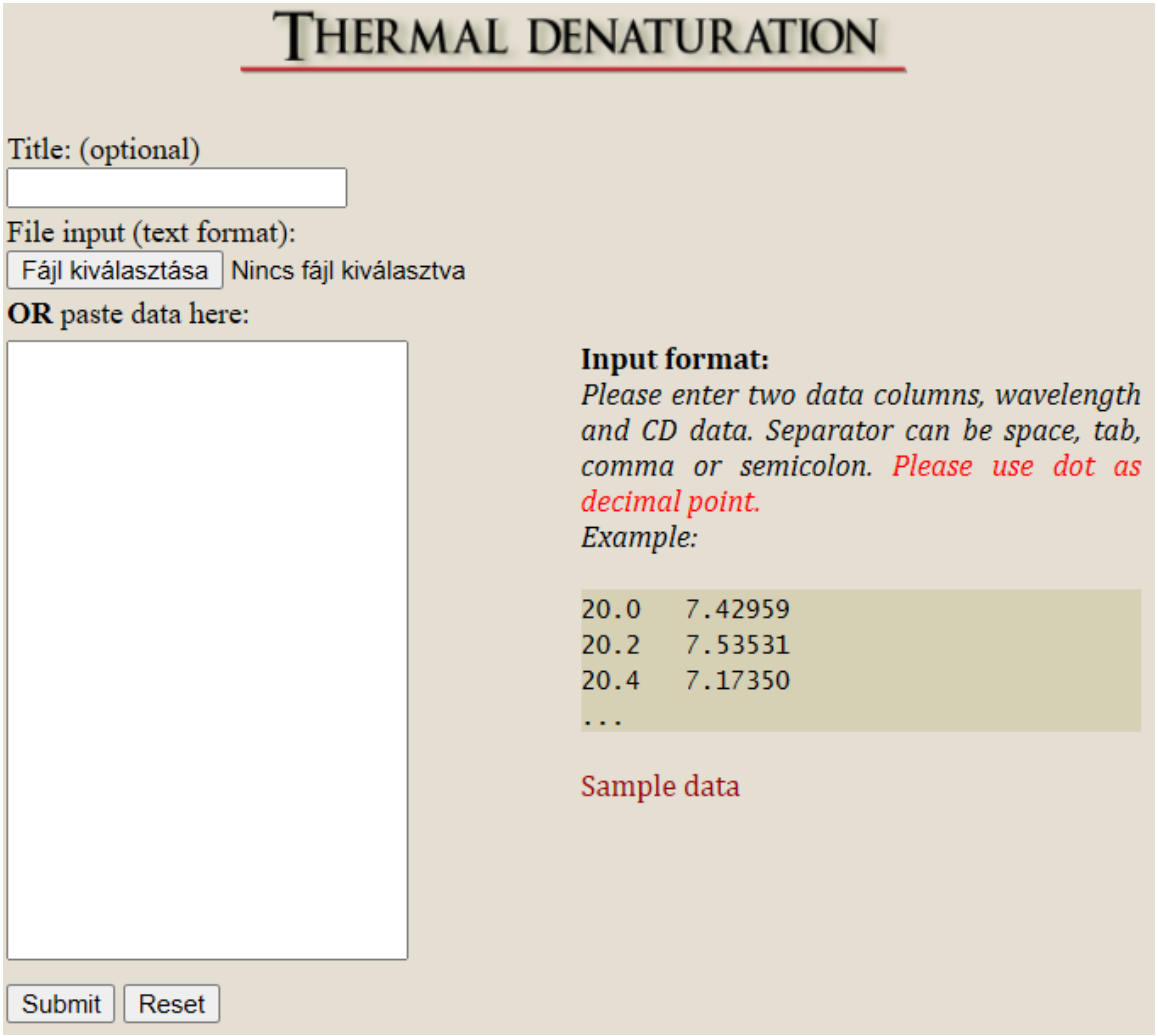
**What are the secondary structure basis components of BeStSel?**

# Thermal denaturation analysis

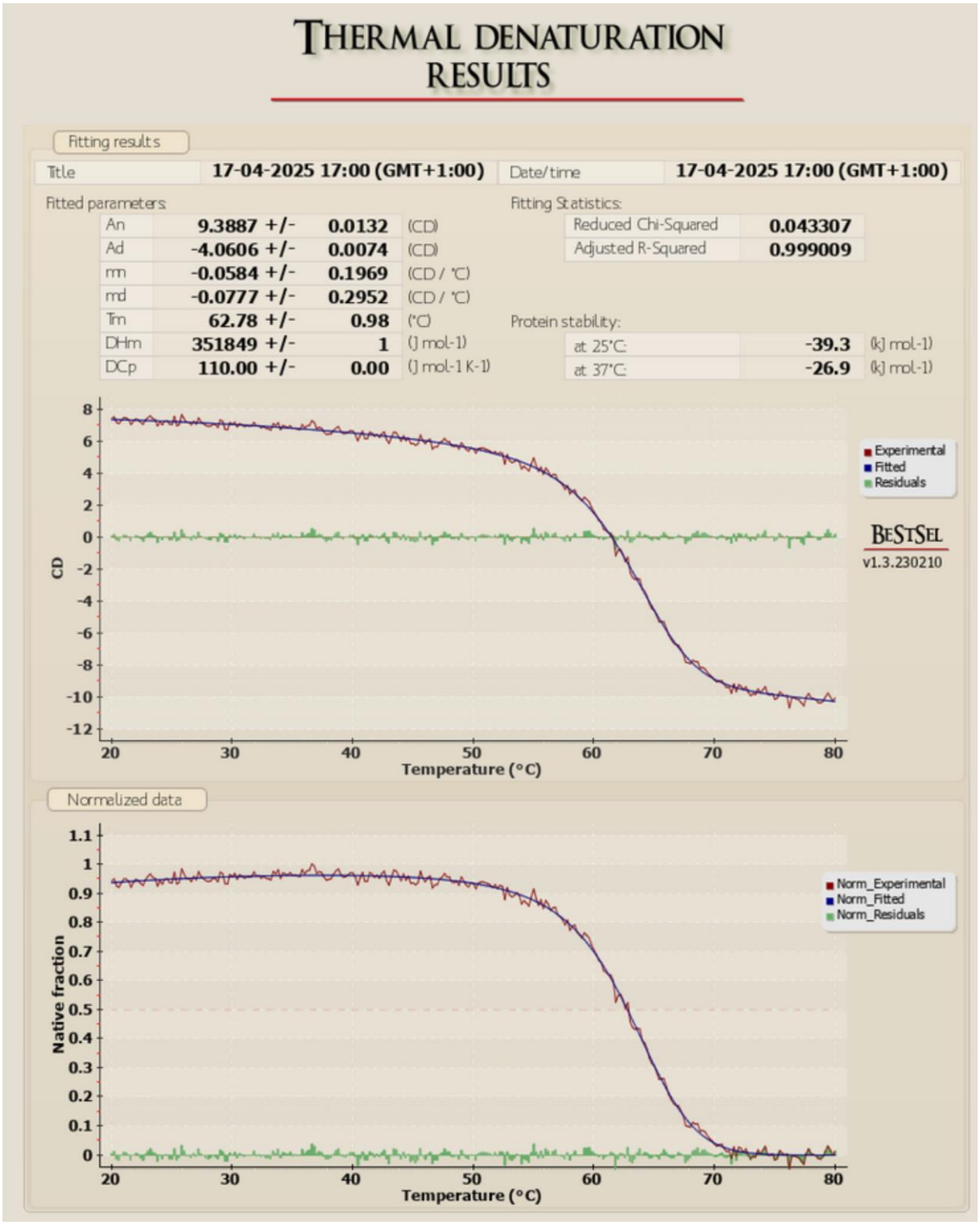
CD spectroscopy is a perfect technique to follow the conformational changes of proteins and thus follow protein unfolding and testing the conformational stability of proteins. The CD spectrum is characteristic to the secondary structure composition and is not sensitive to the environmental factors, such as temperature (provided that the protein structure does not change). This makes CD an excellent technique to follow the thermal stability of proteins vs. fluorescence techniques where the fluorescence efficiency itself is highly temperature dependent. Thermal unfolding profile can be followed by collecting series of spectra or by following the CD signal at a carefully chosen constant wavelength as a function of temperature. Now the webserver supports the analysis of thermal denaturation profiles recorded at a constant wavelength by fitting a two-state model to it, described by Shih et al. (*Protein Sci.* 4, 2050-2062 (1995)).

Data upload

Thermal denaturation profile data in two-column format, temperature and CD data (separators can be space, tab, comma, or semicolon) can be uploaded as text file or can be directly pasted into the window. Using the submit button, data upload and the fitting process is initiated.



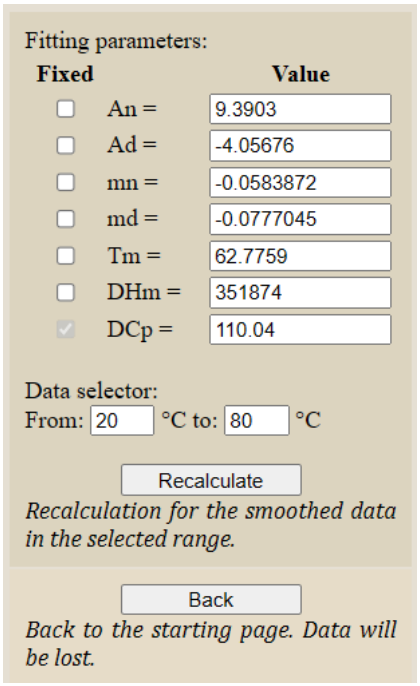
After submission, a result window appears showing the fitted parameters, the raw data and fitted curve at the top, and the calculated “Native fraction” plot below.



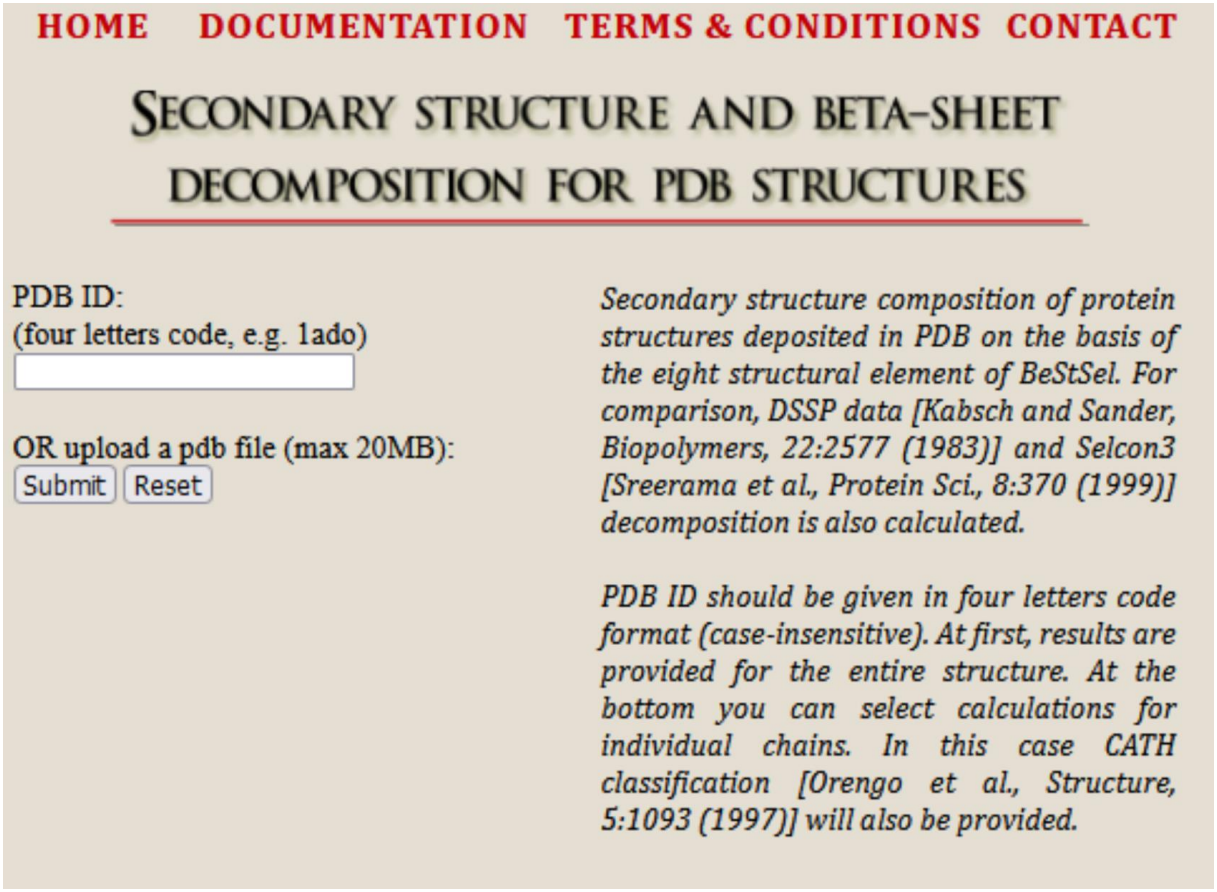
Fitting parameters

The calculation is based on the Gibbs-Helmholtz equation and provides the $T_m$ melting temperature and the $\Delta H_m$ enthalpy change for unfolding at the melting temperature. $A_N$, $A_D$, $m_N$, and $m_D$ are the starting amplitudes of the native and denatured states and their temperature dependence, i.e., the slopes of the denaturation curve before and after the melting transition, respectively. $\Delta C_p$, the heat capacity difference between the denatured and native state is not fitted at the webserver, its value can be set in the fitting function. Providing an experimentally determined or estimated $\Delta C_p$ value, the server calculates the stability of the protein at 25 °C and 37 °C (in kJ/mol). For globular proteins, $\Delta C_p$ is associated with, and can be estimated from the change in the accessible polar and apolar surface areas and in the total buried surface area upon unfolding (Gomez et al. (1995) *Proteins*, 22, 404-412.) (for small globular proteins, $\Delta C_p$ is approximately 50 J/mol/K per residue).

The temperature range can be selected and the fitting recalculated. Fitting parameters can be manually adjusted for recalculation, as well. Note, that the value of $\Delta C_p$ has minor effect on the $T_m$ and $\Delta H_m$ and thus a zero value might be used for an initial fitting.

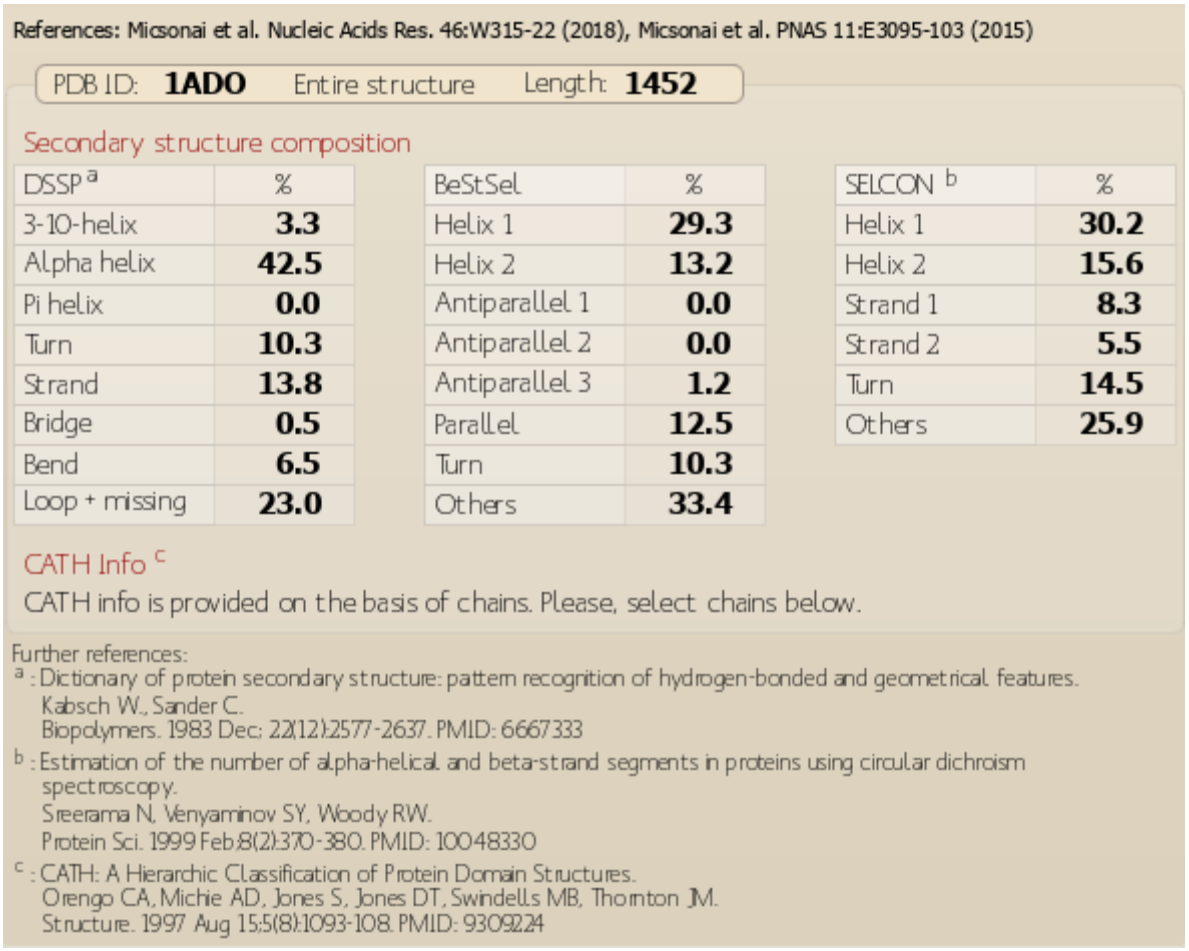


# Secondary structure from PDB files




The „Secondary structure and beta-sheet decomposition for PDB structure" module is used for the calculation of the secondary structure composition of protein structures on the basis of the eight structural elements of BeStSel. For comparison, DSSP data [Kabsch and Sander, Biopolymers, 22:2577 (1983)] and Selcon3 [Sreerama et al., Protein Sci., 8:370 (1999)] composition is also calculated. Either structures deposited in PDB can be submitted or PDB files can be uploaded. In case of submitting a PDB ID, the ID should be given in four letters code format (case-insensitive).

At first, results are provided for the entire structure in the *Result* page of the „Secondary structure from PDB files" module. At the bottom of the page, the labeled polypeptide chains in the structure are listed for selection to display (see below). For the selected individual chains, the CATH classification (Orengo *et al.,* (1997) *Structure* 5(8):1093-1108.) will also be provided (if any).



References: Micsonai et al. Nucleic Acids Res. 46:W315-22 (2018), Micsonai et al. PNAS 11:E3095-103 (2015)

PDB ID: **1ADO** Entire structure Length: **1452**

Secondary structure composition

| DSSP [a] | % |
|---|---|
| 3-10-helix | **3.3** |
| Alpha helix | **42.5** |
| Pi helix | **0.0** |
| Turn | **10.3** |
| Strand | **13.8** |
| Bridge | **0.5** |
| Bend | **6.5** |
| Loop + missing | **23.0** |

| BeStSel | % |
|---|---|
| Helix 1 | **29.3** |
| Helix 2 | **13.2** |
| Antiparallel 1 | **0.0** |
| Antiparallel 2 | **0.0** |
| Antiparallel 3 | **1.2** |
| Parallel | **12.5** |
| Turn | **10.3** |
| Others | **33.4** |

| SELCON [b] | % |
|---|---|
| Helix 1 | **30.2** |
| Helix 2 | **15.6** |
| Strand 1 | **8.3** |
| Strand 2 | **5.5** |
| Turn | **14.5** |
| Others | **25.9** |

CATH Info [c]

CATH info is provided on the basis of chains. Please, select chains below.

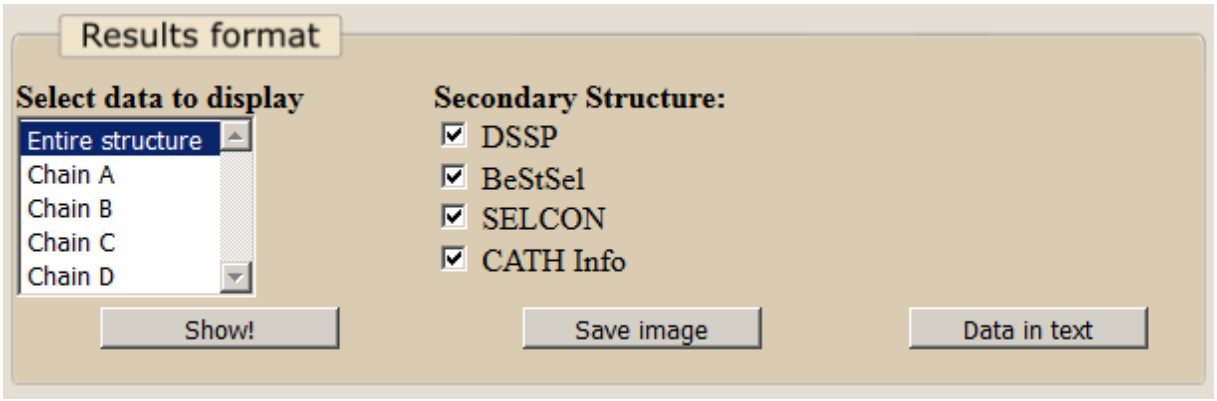
Further references:

[a] : Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.
Kabsch W., Sander C.
Biopolymers. 1983 Dec; 22(12):2577-2637. PMID: 6667333

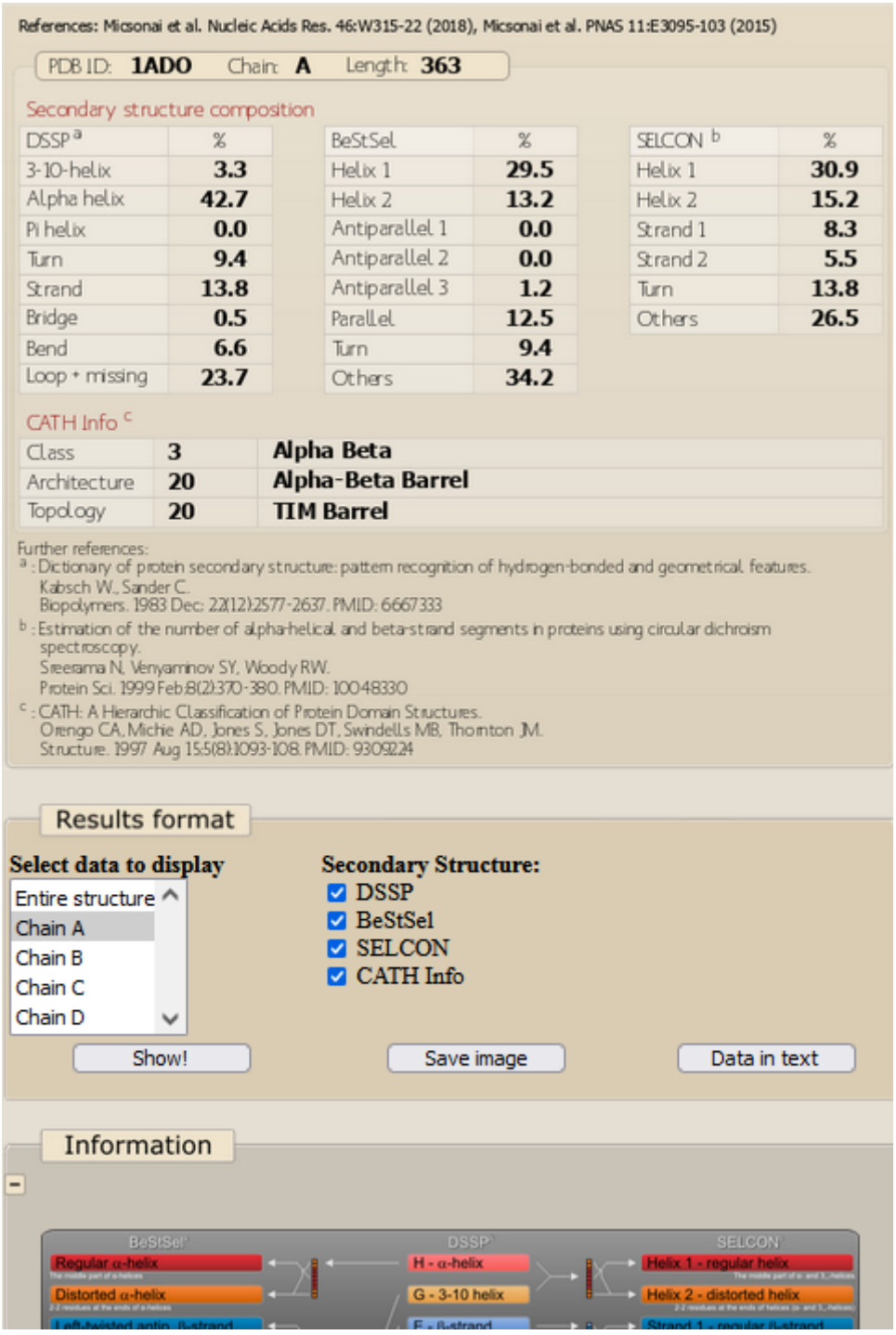
[b] : Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy.
Sreerama N, Venyaminov SY, Woody RW.
Protein Sci. 1999 Feb;8(2):370-380. PMID: 10048330

[c] : CATH: A Hierarchic Classification of Protein Domain Structures.
Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM.
Structure. 1997 Aug 15;5(8):1093-108. PMID: 9309224

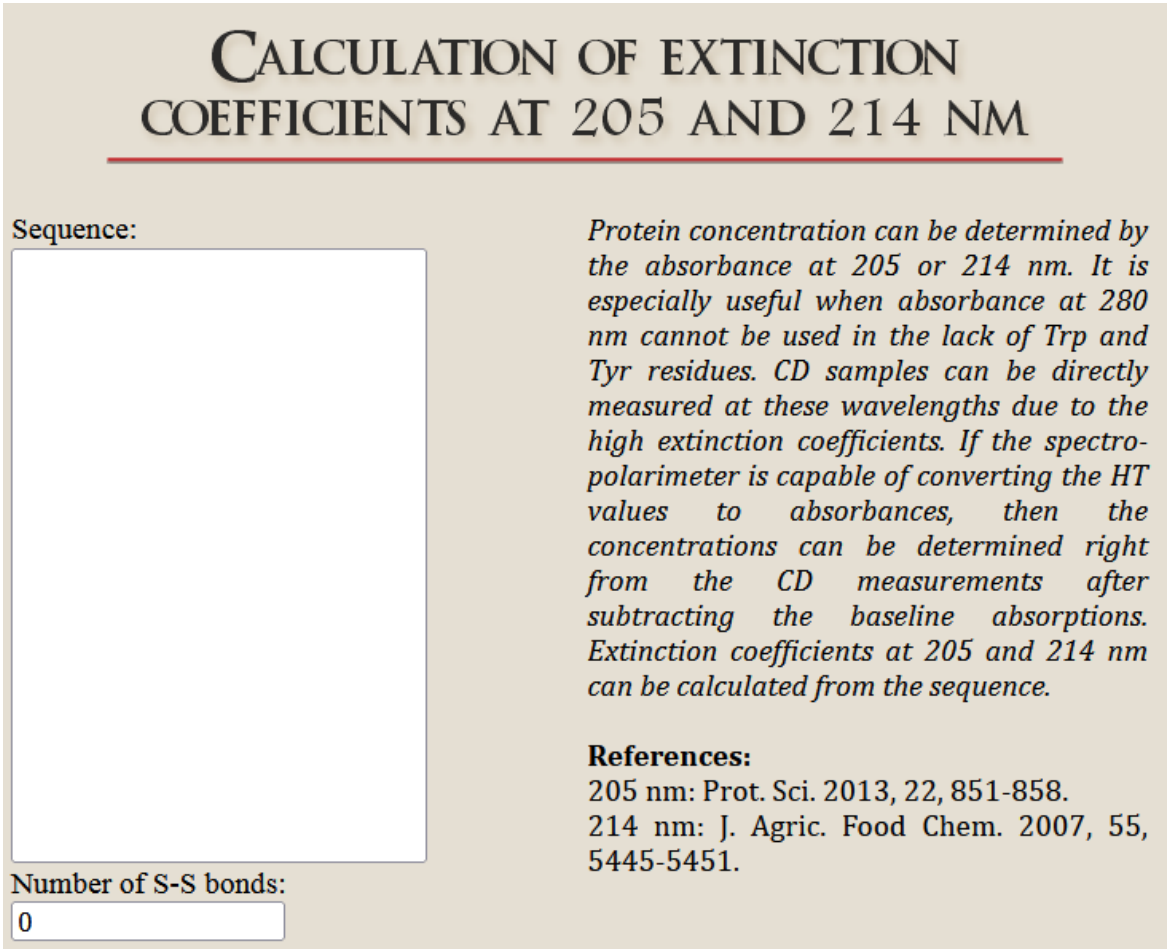
At the bottom of the page the secondary structure decomposition methods can be selected independently to display in a downloadable image or in text format (Data in text).



The secondary structure composition of the entire structure or the selected chains (see below) is displayed separately. The detailed descriptions of the structural elements are described in the original papers, and a brief summary can also be found in the “Information” part of the page.



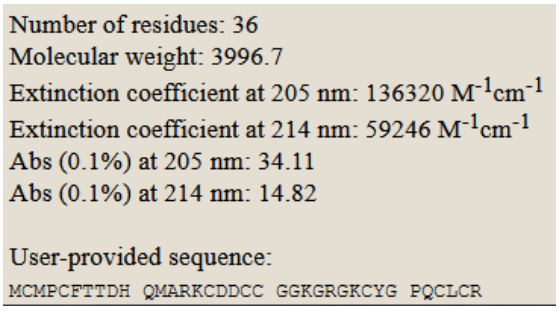
# Extinction coefficient calculator




The “Extinction coefficient calculator” enables the user to determine protein concentrations using absorbance at 205 or 214 nm. This is especially useful when absorbance at 280 nm cannot be used due to the lack of Trp and the small number of Tyr residues.

The absorbance of the CD samples can be directly measured at these wavelengths by the spectropolarimeter. If the instrument is capable of converting HT values to absorbance values, the protein concentration of the sample can be determined from the CD measurement after subtracting the baseline absorptions.

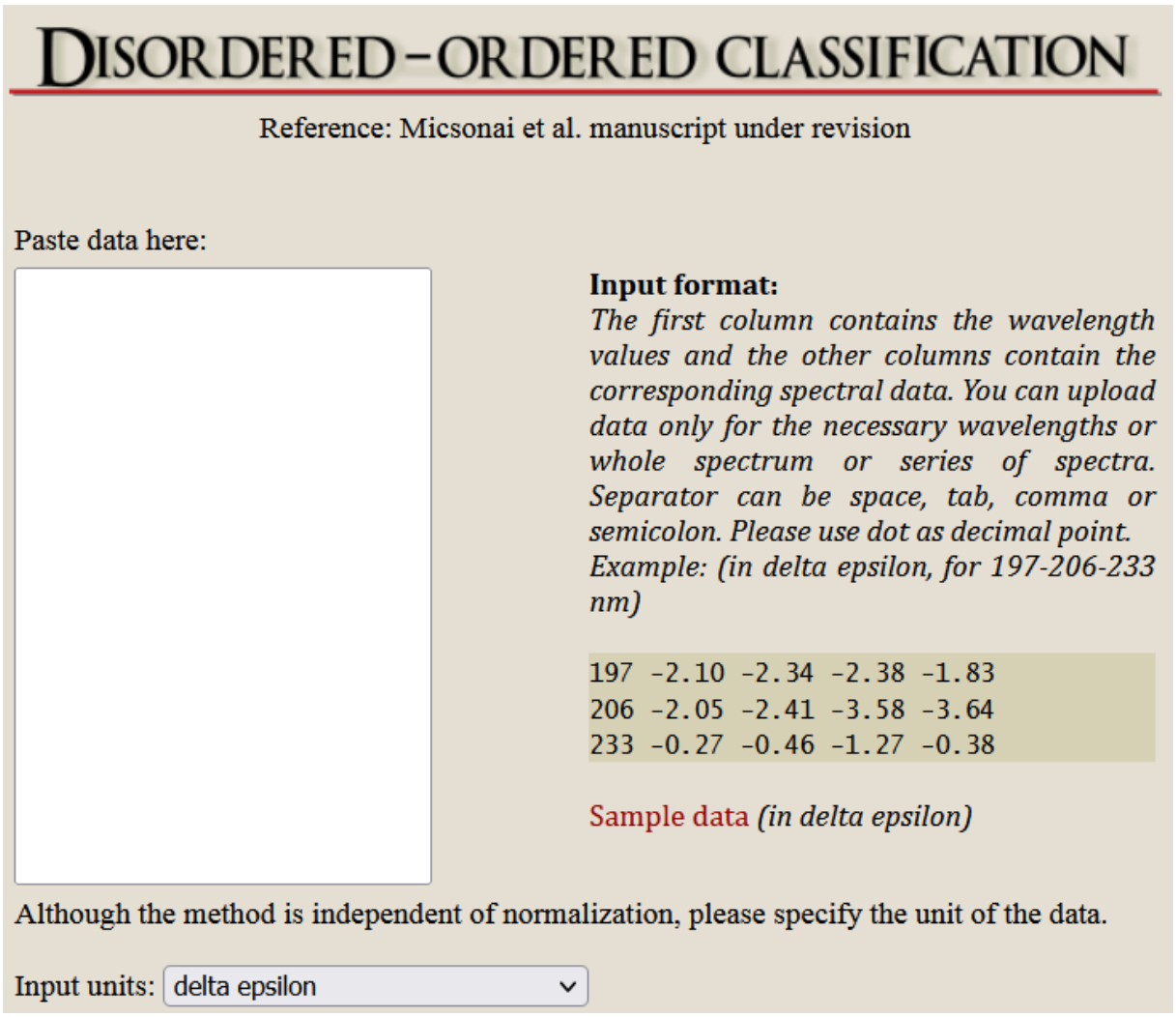
Extinction coefficients at 205 and 214 nm are calculated from the amino acid sequence (see the references for more information) and the number of disulfide bonds. Results are provided at the bottom of the page and also include the number of residues and the molecular weight.



# Disordered-ordered classification

The “Disordered-ordered classification” module categorizes proteins as ordered or disordered based on their CD spectra. This feature is particularly useful to differentiate between disordered and right-hand twisted antiparallel β proteins that have quite similar spectra.

CD data can be provided for a single protein or for multiple proteins. The first column should contain wavelength values and each following column should comprise CD data for a particular protein. Wavelength values must include either 197, 206 and 233 nm or 212, 217 and 225 nm for the predictions.



The results of the classification are provided in a table along with the data used for the predictions.

## DISORDERED–ORDERED CLASSIFICATION

Reference: Micsonai et al. manuscript under revision

| No. | 197 nm | 206 nm | 233 nm | Prediction |
|---|---|---|---|---|
| 1 | 20.20 | 5.36 | -3.48 | ordered |
| 2 | 18.60 | 4.69 | -3.24 | ordered |
| 3 | 17.00 | 4.01 | -3.00 | ordered |
| 4 | 15.40 | 3.34 | -2.76 | ordered |
| 5 | 13.80 | 2.66 | -2.52 | ordered |
| 6 | 12.20 | 1.99 | -2.28 | ordered |
| 7 | 10.60 | 1.31 | -2.04 | ordered |
| 8 | 9.00 | 0.64 | -1.80 | ordered |
| 9 | 7.40 | -0.04 | -1.56 | ordered |
| 10 | 5.80 | -0.71 | -1.32 | ordered |
| 11 | 4.20 | -1.39 | -1.09 | ordered |

# Cited by…

On the “Cited by” page, the user will find a collection of the publications that cited any one of the publications about BeStSel (**Micsonai *et al*. PNAS (2015), Micsonai *et al*. Nucleic Acids Res. (2018), Micsonai *et al*. Methods Mol Biol. (2021), Micsonai et al. Nucleic Acids Res. (2022)**. This page features a search engine to allow users to browse amongst ~2000 articles to find examples and useful information on the applications of CD spectroscopy and BeStSel.

Cited by...

You can browse among ~1000 articles to find examples and useful information on the use of CD spectroscopy and BeStSel
If you would like to be on the list, please cite: Micsonai et al. PNAS (2015) | Micsonai et al. Nucleic Acids Res. (2018) | Micsonai et al. Methods Mol Biol. (2021)

Show 10 entries Search:

| Publication | Title | PubMed | Crossref | Keywords |
|---|---|---|---|---|
| Abelein et al. *J Am Chem Soc* (2016) | Ionic Strength Modulation of the Free Energy Landscape of Aβ40 Peptide Fibril Formation | 27171340 | Crossref | Salts \| Nanofibers \| Peptides and proteins \| Aggregation \| Kinetics |
| Abeyawardhane et al. *ACS Chem Neurosci* (2019) | C-Terminal Cu II Coordination to α-Synuclein Enhances Aggregation | 30384594 | Crossref | EPR spectroscopy \| N-Acetylated α-synuclein \| Parkinson's disease \| aggregation \| metal dyshomeostasis \| missense mutation |
| Abeyawardhane et al. *J Am Chem Soc* (2018) | Copper Induced Radical Dimerization of α-Synuclein Requires Histidine | 30422655 | Crossref | |
| Abeyawardhane et al. *J Am Chem Soc* (2018) | Iron Redox Chemistry Promotes Antiparallel Oligomerization of α-Synuclein | 29608844 | Crossref | |
| Abioye et al. *J Agric Food Chem* (2021) | Inhibition of Islet Amyloid Polypeptide Fibrillation by Structurally Diverse Phenolic Compounds and Fibril Disaggregation Potential of Rutin and Quercetin | 34964624 | Crossref | aggregation \| disaggregation \| fibril formation \| islet amyloid polypeptide \| phenolic compounds \| structure–activity relationship |
| Abouelhadid et al. *mBio* (2020) | Characterization of Posttranslationally Modified Multidrug Efflux Pumps Reveals an Unexpected Link between Glycosylation and Antimicrobial Resistance | 33203757 | Crossref | N-linked glycans \| glycosylation \| multidrug efflux pump |
| Acosta et al. *Biomacromolecules* (2021) | Charge Density as a Molecular Modulator of Nanostructuration in Intrinsically Disordered Protein Polymers | 32840359 | Crossref | |
| Adams et al. *Chem Sci* (2019) | One fold, two functions: cytochrome P460 and cytochrome c'-β from the methanotroph Methylococcus capsulatus (Bath) | 30996884 | Crossref | protein evolution \| methanotroph \| anammox \| metalloproteins \| gas sensors \| hydroxylamine |
| Agadi et al. *J Struct Biol* (2018) | Structural insight into the mechanism of action of antimicrobial peptide BMAP-28(1-18) and its analogue mutBMAP18 | 30336202 | Crossref | |
| Agnolon et al. *Front Immunol* (2020) | Designs and Characterization of Subunit Ebola GP Vaccine Candidates: Implications for Immunogenicity | 33250896 | Crossref | Ebola glycoprotein \| Ebola virus \| glycoprotein \| recombinant vaccine \| trimeric protein \| vaccine. |

Showing 1 to 10 of 991 entries Previous 1 2 3 4 5 … 100 Next

# Documentation

Under the Documentation tab on the front page top right side, various information are available on the BeStSel webserver including this tutorial, the following brief guide to CD spectroscopy measurements and data analysis, and a Q&A section.