



# TUTORIAL

last update: 16-05-2018

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 the CD spectra. These methods often fail to provide acceptable results on  $\alpha/\beta$  mixed or  $\beta$ -structure-rich proteins. The problem arises from the spectral diversity of  $\beta$ -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 the  $\beta$ -sheets account for the observed spectral diversity. We developed the Beta Structure Selection method (BeStSel) for the secondary structure estimation that takes into account the twist of  $\beta$ -structures. This method can reliably distinguish parallel and antiparallel  $\beta$ -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 CA, *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.


Here, we provide a brief introduction for the use of the BeStSel web server <http://bestsel.elte.hu>. The server is under 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](mailto:kardos@elte.hu) or [micsonai@ttk.elte.hu](mailto: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 publication of Micsonai *et al.*

<b>Introduction</b>	<b>Page 2</b>
<b>Single spectrum analysis module</b>	<b>Page 3</b>
<b>Multiple spectra analysis module</b>	<b>Page 16</b>
<b>Secondary structure and beta-sheet decomposition for PDB structure module</b>	<b>Page 20</b>
<b>Fold recognition module</b>	<b>Page 23</b>

# Introduction


First, one of the 4 modules of the server can be chosen, listed on the left side of the starting page, *Single spectrum analysis*, *Fold recognition*, *Multiple spectra analysis*, and *Secondary structure and beta-sheet decomposition for PDB structures*. In *Single spectrum analysis*, a single CD spectrum can be analyzed for the secondary structure composition and the protein fold can be predicted.

**BeStSel** TEST VERSION


BeStSel (Beta Structure Selection) is a novel method for the secondary structure determination and fold recognition from protein circular dichroism spectra.

BeStSel was tested and optimized for Firefox. At the moment, it works defectively under Windows Explorer.

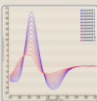
**Single spectrum analysis**  
Secondary structure determination distinguishing parallel beta-sheets and antiparallel beta-sheets of different twists, and fold recognition from the CD spectrum.




**Fold recognition**  
Prediction of fold class, architecture, topology and homology for the provided secondary structure contents.



**Multiple spectra analysis**  
Analysis of a series of spectra as a function of temperature, time, ligand concentration, etc.



**Secondary structure and beta-sheet decomposition for PDB structures**  
Secondary structure composition of protein structures deposited in PDB on the basis of the eight structural element decomposed by BeStSel. For comparison, DSSP data and Selcon3 decomposition is also calculated.



HOME INFORMATION TERMS & CONDITIONS CONTACT

## SINGLE SPECTRUM ANALYSIS & FOLD RECOGNITION

Reference: Micsonai et al. PNAS 112:E3095-103 (2015).

Title: (optional)

File input (text format):  
 Nincs kijelölve fájl.

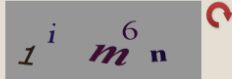
OR paste data here:

**Input format:**  
Please enter two data columns, wavelength and CD data. Separator can be space, tab, comma or semicolon. Please use dot as decimal point.  
Example:





175	-0.47423
176	-0.35859
177	-0.25112
...	

Sample spectrum (in delta epsilon)

Input units:



Please, type the characters above or use your password.



## Single spectrum analysis

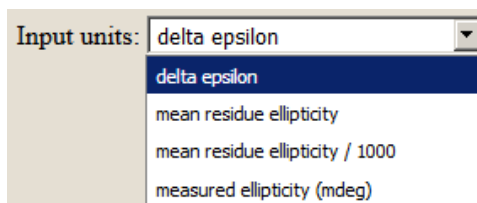
Data can be uploaded from a text file or can be copied into the window in two data columns, separator can be space, tab, comma or semicolon. Please use dot as decimal point. In case of browsed data file in text format, the system automatically recognize the header and the data columns.

### Input units

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

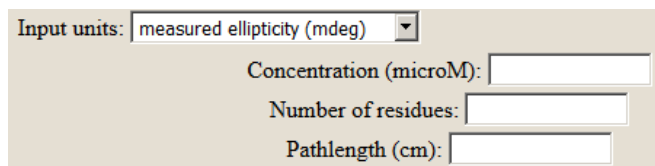
**Delta epsilon** ( $\text{M}^{-1} \text{cm}^{-1}$ )

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



Input units: delta epsilon  
delta epsilon  
mean residue ellipticity  
mean residue ellipticity / 1000  
measured ellipticity (mdeg)

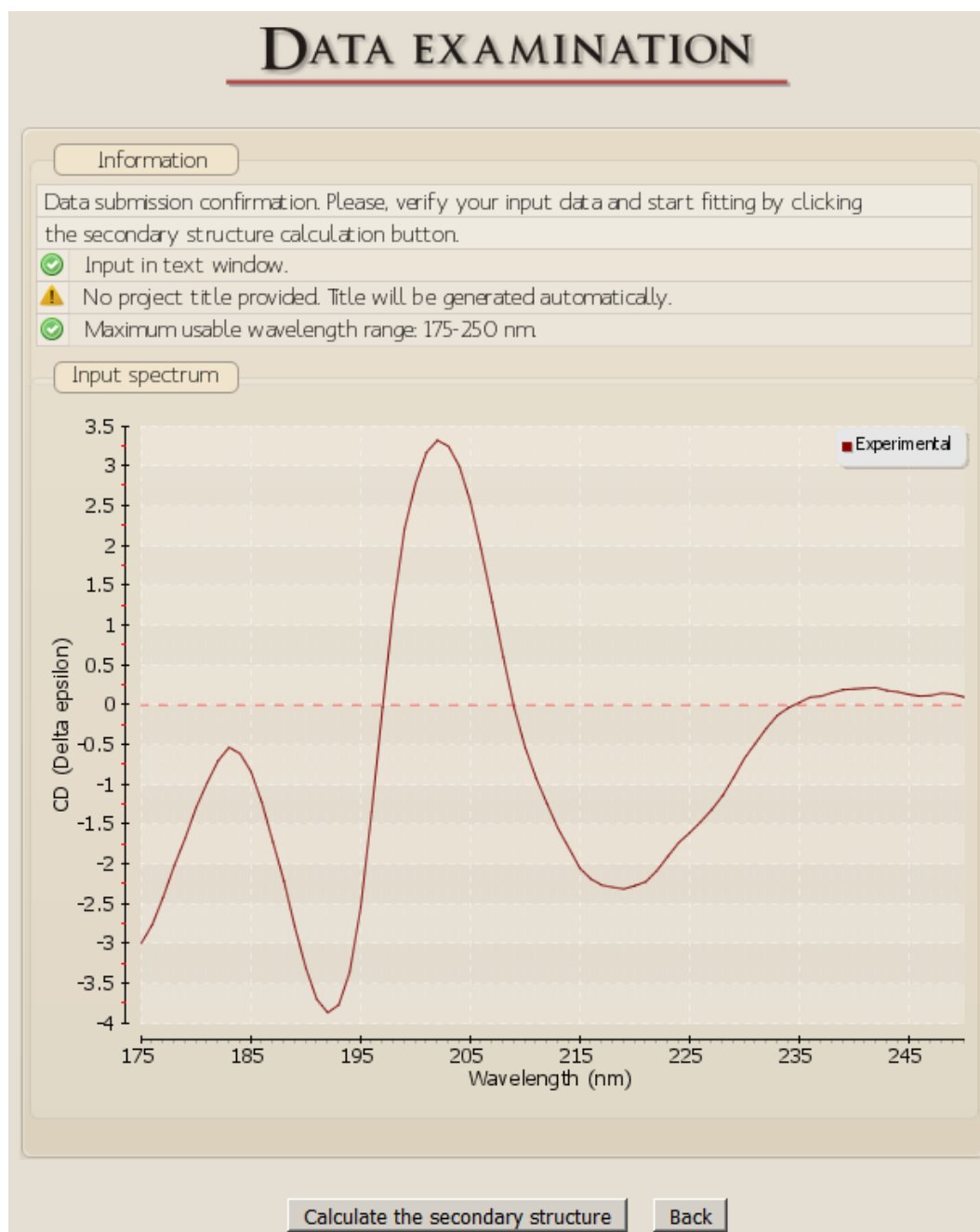
**Measured ellipticity** data can be directly uploaded. In that case, the protein concentration in  $\mu\text{M}$ , the number of residues per protein molecule and the pathlength in cm should be provided by the user.



Input units: measured ellipticity (mdeg)  
Concentration (microM):   
Number of residues:   
Pathlength (cm):

At the bottom, please, provide the captcha or use a password to submit your data. This is only to avoid the attack of robots, there is no need for registration to use the server.

A *Data examination* window will appear to check if the data was uploaded properly. Data is converted to  $\Delta\epsilon$ .



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. The system makes an automatic data examination and gives a message in case of unexpected CD amplitudes (calculation is still possible).

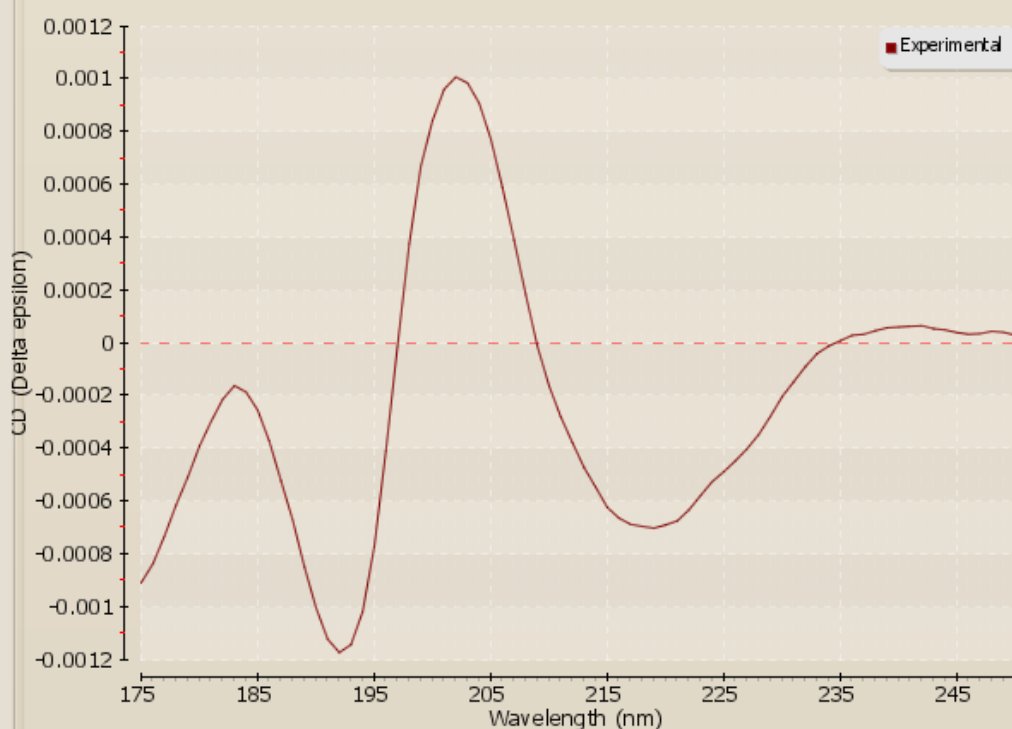
# DATA EXAMINATION

## Information

Data submission confirmation. Please, verify your input data and start fitting by clicking the secondary structure calculation button.

- ✓ Input in text window.
- ⚠ CD units are converted to delta epsilon.
- ⚠ Amplitude is too low. Please, check units and normalization.
- ⚠ No project title provided. Title will be generated automatically.
- ✓ Maximum usable wavelength range: 175-250 nm.

## Input spectrum



Possible normalization or unit problem, anyway, calculate the secondary structure

Back

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. 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 uses 8 precalculated and fixed basis spectra sets - which are optimized for the chosen wavelength range - to analyze the submitted spectrum and estimate the secondary structure content. For all details on the optimization and fitting processes, refer to the original publication of Micsonai et al.

# SINGLE SPECTRUM ANALYSIS RESULTS

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

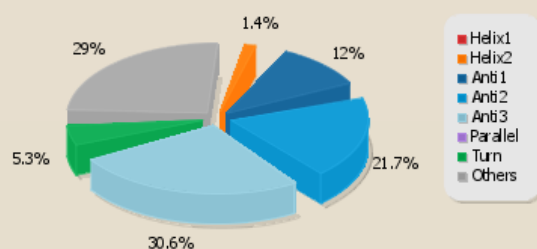
## Header

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

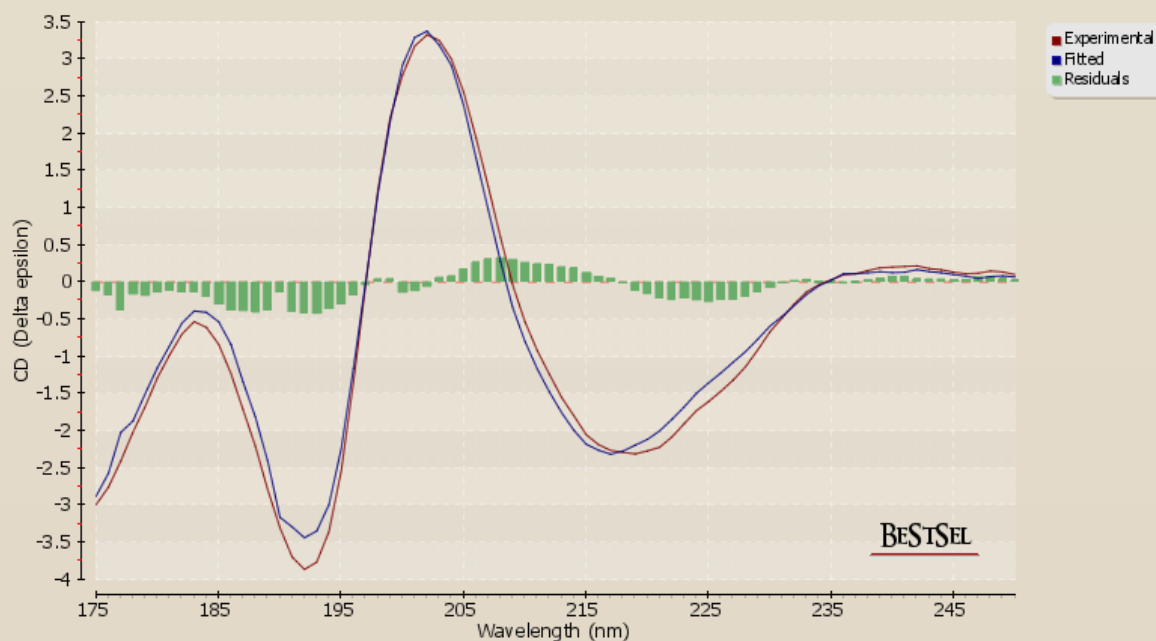
## Secondary structure

Estimated secondary structure content (%)

Helix	1.4	Helix1 (regular)	0.0
		Helix2 (distorted)	1.4
Antiparallel	64.2	Anti1 (left-twisted)	12.0
		Anti2 (relaxed)	21.7
		Anti3 (right-twisted)	30.6
Parallel	0.0		
Turn	5.3		
Others	29.0		



## Fitting results



RMSD **0.02029**

NRMSD **0.02822**

## Comments

- ✓ Input in text window.
- ⚠ No project title provided. Title will be generated automatically.
- ✓ Maximum usable wavelength range: 175-250 nm

## 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.

The 'Results format' dialog box is shown with the following options:

Header	Secondary structure	Plot	Comments
<input checked="" type="checkbox"/> Title	<input checked="" type="checkbox"/> 8 structures	<input checked="" type="checkbox"/> Plot	<input checked="" type="checkbox"/> Warnings/Input details
<input checked="" type="checkbox"/> Date/time	<input type="radio"/> No pie	<input checked="" type="checkbox"/> RMSD	<input type="checkbox"/> User comments
<input checked="" type="checkbox"/> Wavelength range	<input type="radio"/> HAPTO pie	<input checked="" type="checkbox"/> NRMSD	
<input checked="" type="checkbox"/> Scale factor	<input checked="" type="radio"/> 8 structures pie		

User comments:

Buttons: Show! Save image Data in text Make PDF

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.  $\sqrt{\frac{1}{w} \sum_{i=1}^w (CD_{exp,i} - CD_{fit,i})^2}$

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

## Data in text

For further data procession 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.

```
Data in text

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

Header
=====
Wavelength range: 175-250 nm
Scale factor: 1

Estimated secondary structure content (%)
=====
Helix1 (regular): 0.0
Helix2 (distorted): 1.4
Anti1 (left-twisted): 12.0
Anti2 (relaxed): 21.7
Anti3 (right-twisted): 30.6
Parallel: 0.0
Turn: 5.3
Others: 29.0
-----
Helix: 1.4
Antiparallel: 64.2
Parallel: 0.0
Turn: 5.3
Others: 29.0

Spectral deviation
=====
RMSD: 0.2029
NRMSD: 0.02822

Fitting results
=====
Data units: Delta epsilon

Wavelength (nm) Experimental Fitted Residuals
175 -2.9917 -2.8767 -0.115
176 -2.7581 -2.5712 -0.1869
177 -2.408 -2.0231 -0.3849
178 -2.0188 -1.8683 -0.1505
179 -1.6721 -1.4994 -0.1727
180 -1.2863 -1.1492 -0.1371
181 -0.9769 -0.8522 -0.1247
182 -0.6991 -0.5513 -0.1478
183 -0.5313 -0.3886 -0.1428
184 -0.6107 -0.4074 -0.2033
185 -0.8382 -0.534 -0.3042
186 -1.2299 -0.839 -0.3909
187 -1.7227 -1.3495 -0.3732
188 -2.2111 -1.818 -0.3931
189 -2.7929 -2.4095 -0.3834
190 -3.3049 -3.1639 -0.141
191 -3.6993 -3.294 -0.4053
192 -3.8653 -3.439 -0.4263
193 -3.7669 -3.3487 -0.4182
194 -3.3544 -2.993 -0.3614
```



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

**Information**

**How the fitting to the CD spectrum is carried out in BeStSel?**

BeStSel fits the experimental CD curve by the linear combination of fixed basis components to get the proportion of the eight structural elements.

**Considerations for wavelength range selection and the importance of the correct protein concentration**

BeStSel automatically offers the available wavelength ranges for calculations by inspecting the uploaded data. Choose the widest wavelength range for which the absorption (HT) is within the acceptable limit (cut-off). The fitting reliability strongly depends on the correct protein concentration and cell pathlength, chosen to normalize the CD spectra. In case of uncertainty in the concentration and pathlength measures, a best correction factor may be calculated by BeStSel, providing the lowest rmsd. The calculation is efficient when using the wide wavelength range (175-250 nm). In case of correct normalization, fittings for different wavelength ranges should provide similar results.

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

The secondary structure basis components of BeStSel are derived from DSSP. We introduced novel subgroups of the beta-sheets based on the beta-sheet twist. Parallel and antiparallel beta-sheets are distinguished and antiparallel beta-sheets are divided into three subgroups: left-hand twisted, relaxed, and right-hand twisted (anti1, anti2, anti3, respectively). The regular part of helices (helix1) and the distorted ends (helix2) are separated, similarly to SELCON3, however, only  $\alpha$ -helices are counted. BeStSel sorts  $3_{10}$ -helix to "others". The definition of turn is identical to that in DSSP. The figure shows the eight basis components of BeStSel in relation to DSSP. For comparison, basis components of SELCON3 algorithm, which are also used for CONTIN and CDSSTR in CDPro, are presented.

<sup>1</sup> Mousset et al. (Manuscript in preparation)

<sup>2</sup> Dictionary of protein secondary structure: pattern-recognition of hydrogen-bonded and geometrical features. Kabsch, U., Sander, C. Biopolymers, 1983 Dec; 23(12):3577-3632.

Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism. Jemnitz, M., Vothkamm, S., Bressan, S. Protein Sci. 1998 Feb; 7(2):351-360.

### Wavelength range, scale factor, best factor

Choose wavelength range:

☒ 175-250 nm

☐ 180-250 nm

☐ 190-250 nm

☐ 200-250 nm

Scale factor

*Recalculation for the spectrum multiplied by the scale factor in the chosen wavelength range.*

*Dependence of the secondary structure estimation and NRMSD of fitting on the spectral amplitude (use of the latter as hint for correction is advised only for the 175-250 and 180-250 nm ranges).*

*Prediction of fold class, architecture and topology (CATH classification) based on the results of the secondary structure determination. This function is available when the secondary structure is estimated in the Single spectrum analysis.*

*Back to the starting page. Data will be lost.*

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 recalculation, as well. The CD amplitude is multiplied with this factor.

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

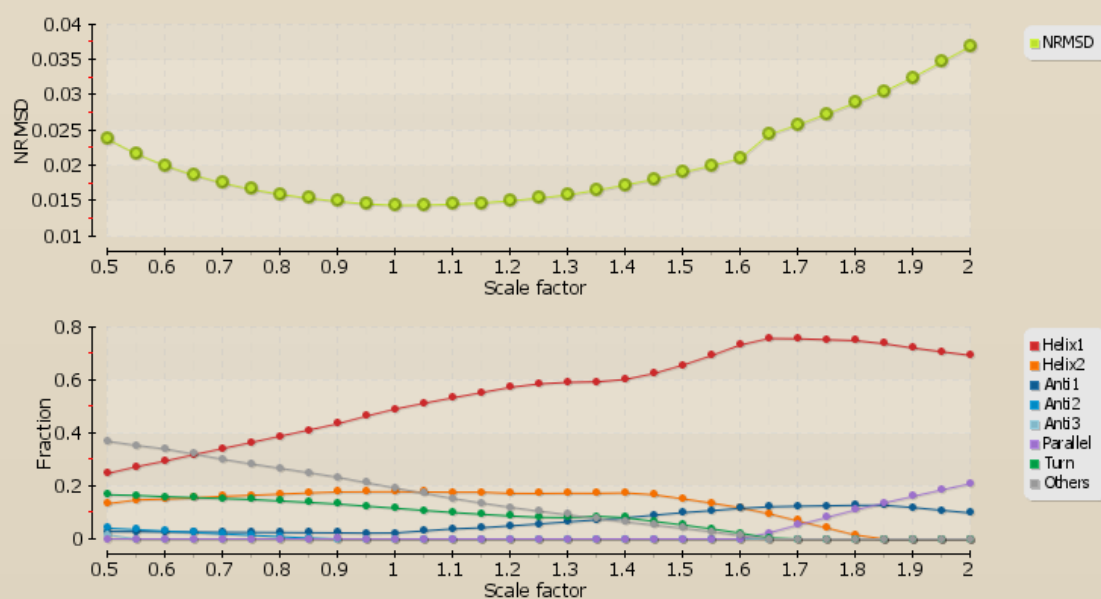
The factor related to the lowest NRMSD is highlighted (see next page). The dependence of the individual secondary structure components on the CD amplitude is plotted. This can be informative in the case of uncertainties in the protein concentration or pathlength. In case of CD data in a wide wavelength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

Here, we show the example of *CIC chloride channel* protein from the Protein Circular Dichroism Data Bank (Whitmore et al., Nucleic Acids Research, 45(D1):D303-D307 (2017), PCDDDB ID: CD0000104000), which has an X-ray structure (PDB:1KPK, Dutzler et al., Nature, 415:287-294 (2002)). Analyses of the spectrum in all the available wavelength ranges (180-250, 190-250 and 200-250 nm) provide acceptable results. We show the dependence of the structure prediction on the spectral amplitude in the different wavelength ranges below. The “Best factor” of the lowest NRMSD works for the 180-250 nm wavelength range, correctly suggesting a factor of 1.0 here. However, for the 190-250 and 200-250 nm ranges, only the amplitude dependence is meaningful, rescaling with the “Best factor” mess up the secondary structure results and may lead erroneous data interpretations.

## Header

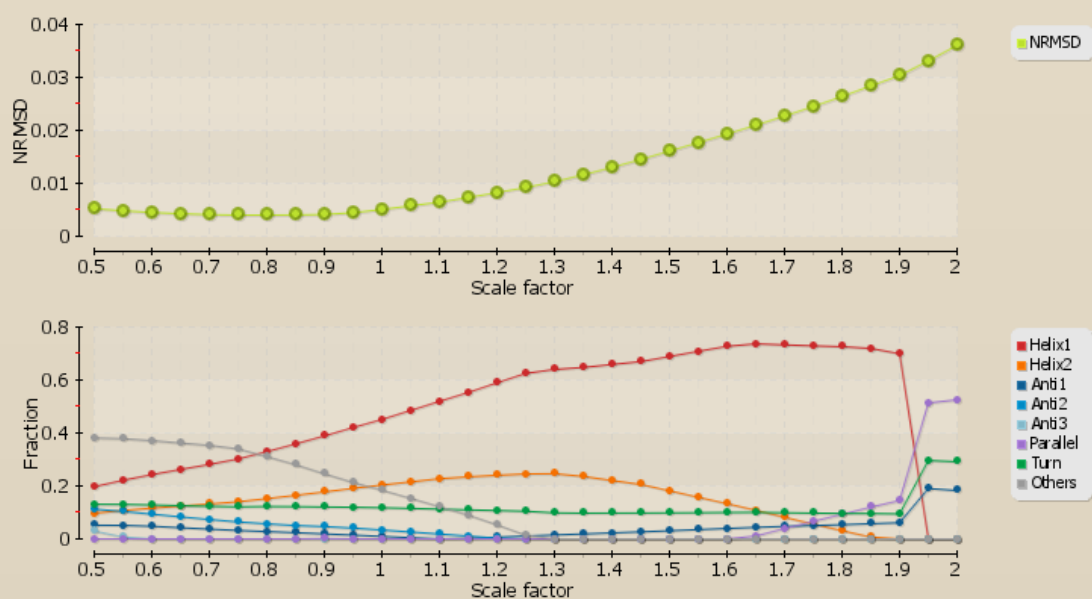
Title	CIC-ec1 CD0000104000	Wavelength range	180-250 nm
Date/time		Scale factor	1



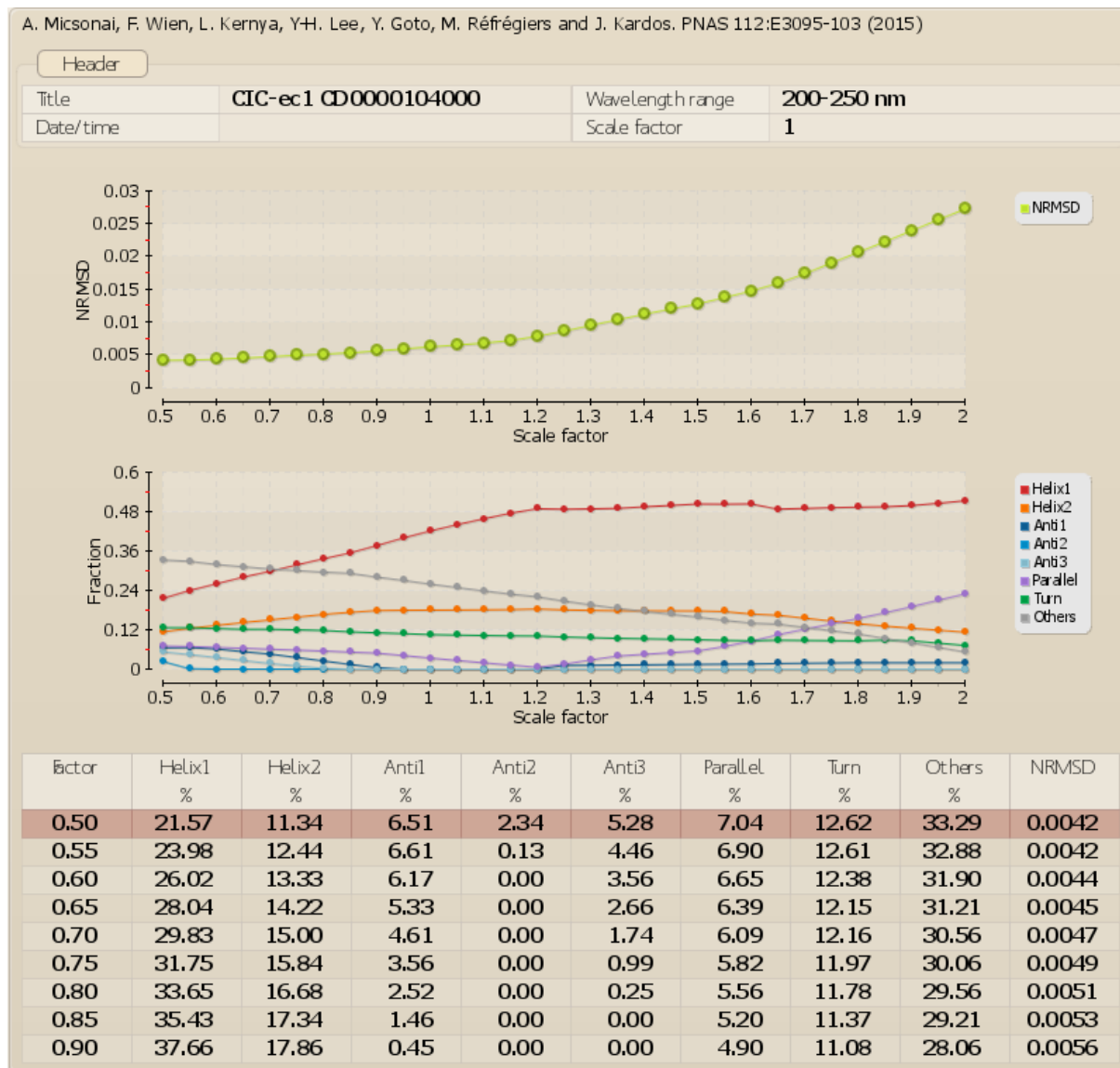
Factor	Helix1 %	Helix2 %	Anti1 %	Anti2 %	Anti3 %	Parallel %	Turn %	Others %	NRMSE
0.80	38.76	16.92	2.42	0.85	0.00	0.00	14.34	26.70	0.0159
0.85	41.10	17.36	2.37	0.33	0.00	0.00	13.79	25.06	0.0154
0.90	43.46	17.72	2.30	0.00	0.00	0.00	13.18	23.33	0.0149
0.95	46.34	17.79	2.20	0.00	0.00	0.00	12.39	21.28	0.0146
1.00	49.03	17.83	2.26	0.00	0.00	0.00	11.60	19.28	0.0144
1.05	51.21	17.74	3.04	0.00	0.00	0.00	10.78	17.25	0.0144
1.10	53.26	17.64	3.77	0.00	0.00	0.00	10.00	15.33	0.0145
1.15	55.23	17.66	4.21	0.00	0.00	0.00	9.31	13.59	0.0147
1.20	57.22	17.20	4.92	0.00	0.00	0.00	8.68	11.97	0.0150

## Header

Title	CIC-ec1 CD0000104000	Wavelength range	190-250 nm
Date/time		Scale factor	1



Factor	Helix1 %	Helix2 %	Anti1 %	Anti2 %	Anti3 %	Parallel %	Turn %	Others %	NRMDS
0.60	24.28	11.62	4.78	9.37	0.00	0.00	12.83	37.12	0.0045
0.65	26.24	12.41	4.24	8.34	0.00	0.00	12.56	36.22	0.0043
0.70	28.14	13.18	3.70	7.34	0.00	0.00	12.30	35.34	0.0041
0.75	30.19	14.04	3.21	6.42	0.00	0.00	12.13	34.01	0.0040
0.80	32.96	15.21	2.79	5.71	0.00	0.00	12.24	31.09	0.0040
0.85	35.88	16.46	2.36	5.07	0.00	0.00	12.19	28.04	0.0040
0.90	38.89	17.75	1.91	4.67	0.00	0.12	12.07	24.59	0.0041
0.95	42.06	19.10	1.45	4.10	0.00	0.00	11.96	21.33	0.0045
1.00	45.12	20.37	0.96	3.38	0.00	0.00	11.77	18.40	0.0051



	Factor	WL range	Helix1	Helix2	Anti1	Anti2	Anti3	Parallel	Turn	Others	NRMSE
X-ray			46.4	16.1	0.0	0.0	0.0	0.0	12.1	25.4	
	Factor										
	1.00	180-250	49.0	17.8	2.3	0.0	0.0	0.0	11.6	19.3	0.0144
	1.00	190-250	45.1	20.4	1.0	3.4	0.0	0.0	11.8	18.4	0.0051
lowest NRMSE	1.00	200-250	42.1	18.0	0.0	0.0	0.0	3.4	10.5	26.0	0.0062
	1.00	180-250	49.0	17.8	2.3	0.0	0.0	0.0	11.6	19.3	0.0144
	<del>0.80</del>	<del>190-250</del>	<del>33.8</del>	<del>15.2</del>	<del>2.8</del>	<del>5.7</del>	<del>0.0</del>	<del>0.0</del>	<del>12.2</del>	<del>31.1</del>	<del>0.0040</del>
	<del>0.50</del>	<del>200-250</del>	<del>21.6</del>	<del>11.3</del>	<del>6.5</del>	<del>2.3</del>	<del>5.3</del>	<del>7.0</del>	<del>12.6</del>	<del>33.3</del>	<del>0.0042</del>

Table: BeStSel analysis of CIC chloride channel and correction with the “best factors” which is not advised for 190-200 and 200-250 nm wavelength ranges.

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.

## Fold recognition

Protein fold can be predicted from the results of the CD spectrum analysis.

### Fold recognition results

At first, 3 different analyses are provided, a search for similar structures on the entire PDB, a fold search on the closest structures on a non-redundant single domain PDB subset, and a search on single domains with secondary structure composition within the expected error of the secondary structure analysis. After this three, by providing the chain length, a fourth, more sophisticated fold prediction can be carried out which uses the weighted K-nearest neighbors search method.

FOLD RECOGNITION RESULTS						
<b>Closest structures in the entire PDB</b>						
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) 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.						

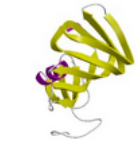


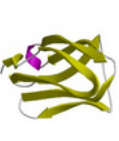


<b>Closest single CATH domains in a PDB subset filtered for &lt;= 90% sequence homology.</b>								
The ten closest structures based on the euclidean distance in the eight dimensional space of BeStSel.								
No.	PDB	Chain	C	A	T	Class	Architecture	Topology
1	1VTZ	2	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
2	1J58	A	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
3	1JHJ	A	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
4	1G6E	A	2	60	20	Mainly Beta	Sandwich	Gamma-B Crystallin; domain 1
5	2PHL	A	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
6	2W2A	A	2	40	128	Mainly Beta	Beta Barrel	Lipocalin
7	1LED	A	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
8	1ZKO	A	2	40	50	Mainly Beta	Beta Barrel	OB fold (Dihydrolipoamide Acetyltransferase, E2P)
9	1QWR	A	2	60	120	Mainly Beta	Sandwich	Jelly Rolls
10	1B35	B	2	60	120	Mainly Beta	Sandwich	Jelly Rolls

### Structural domains within the expected error of BeStSel

The algorithm searches on a PDB subset with  $\leq 90\%$  sequential homology for all the chains having single CATH domains that lie within  $1.5 \times \text{RMSD}$  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
					

In the case of the weighted K- nearest neighbors method the number of residues is required for the analysis.

### Weighted K-nearest neighbors search

Chain length:



Class			
No.	CATH	Category name	Score
1	2	Mainly Beta	17.34
Architecture			
No.	CATH	Category 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	Category 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	Category 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 4.2 (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 WKNN scores calculated excluding every structure belongs to an already predicted categories (lower numbered hits). The WKNN score defined by the sum of the weighted distance of every structures (from the query point) among the K- nearest neighbors which belong to the certain category.

Number of	CATH 4.2
Domains	55350
Class	4
Architecture	41
Topology	1310
Homology	5398



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 structure 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  $i^{\text{th}}$  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  $\leq 90\%$  sequence homology and resolution better than 3.0 Angstr  $\text{\AA}$ 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 account 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  $\leq 90\%$  sequential homology for all the chains having single CATH domains that lie within  $1.5 \times \text{RMSD}$  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 frequencies 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?

## 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 variable as the function of which the spectra were recorded. 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.

### MULTIPLE SPECTRA ANALYSIS

Title: (optional)

File input (text format):  
 Nincs kijelölve fájl.

OR paste data here:

**Input format:**  
*The first row should contain the values of the variable as the function of which the spectra were recorded. 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. Example:*

25	30	35	40	
175	-5.29040	-5.0973	-4.9042	-4.7110
176	-5.68400	-5.4216	-5.1593	-4.8969
177	-5.95660	-5.6231	-5.2895	-4.9560
...				

**Sample spectrum (in delta epsilon)**

Data as a function of:

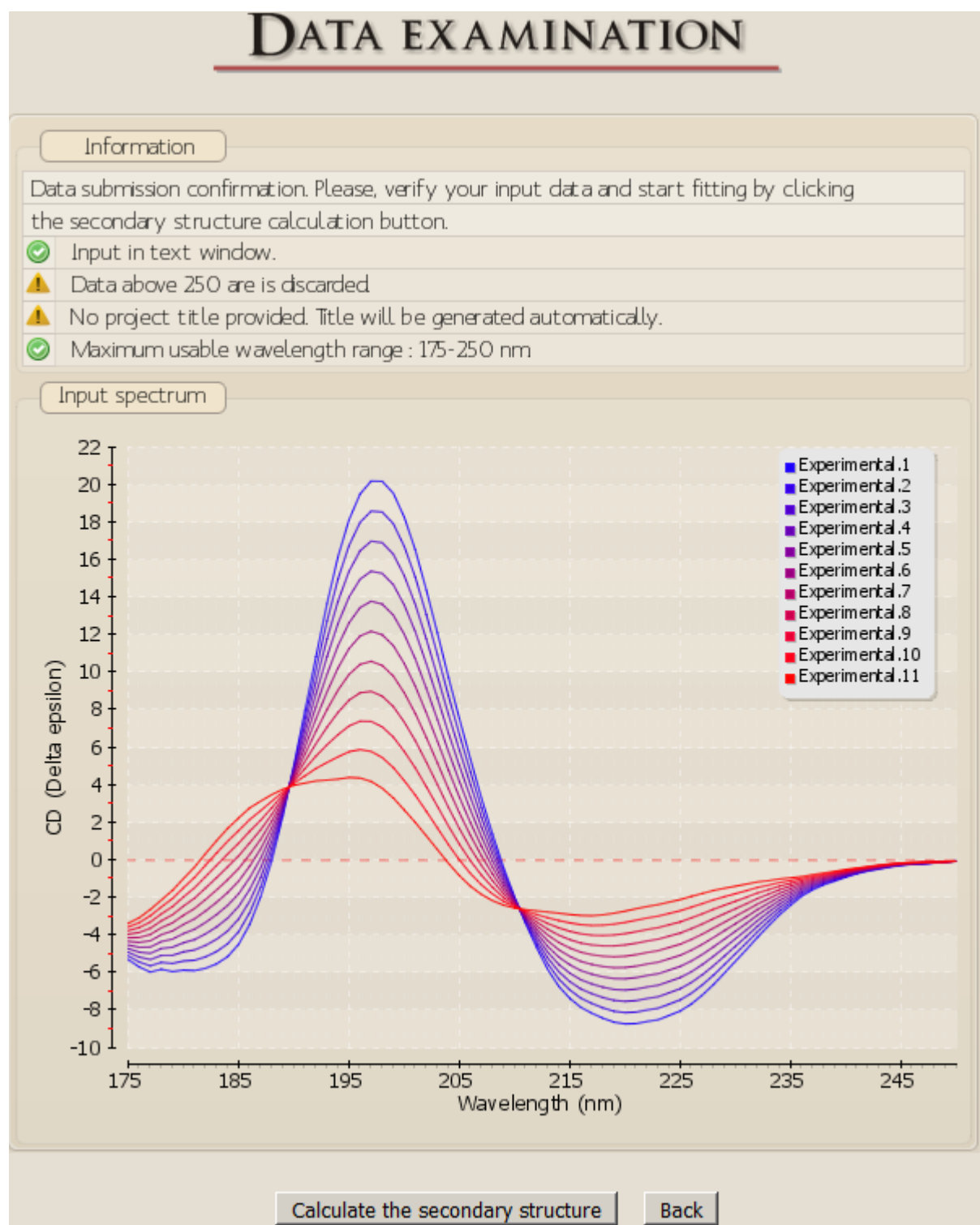
unit:

Input units:

Data as a function of:

Input units:

First a data examination page comes up to check the correct upload. Then, all the spectra will be evaluated at once and shown as a function of the chosen parameter.



Input spectrum

CD (Delta epsilon)

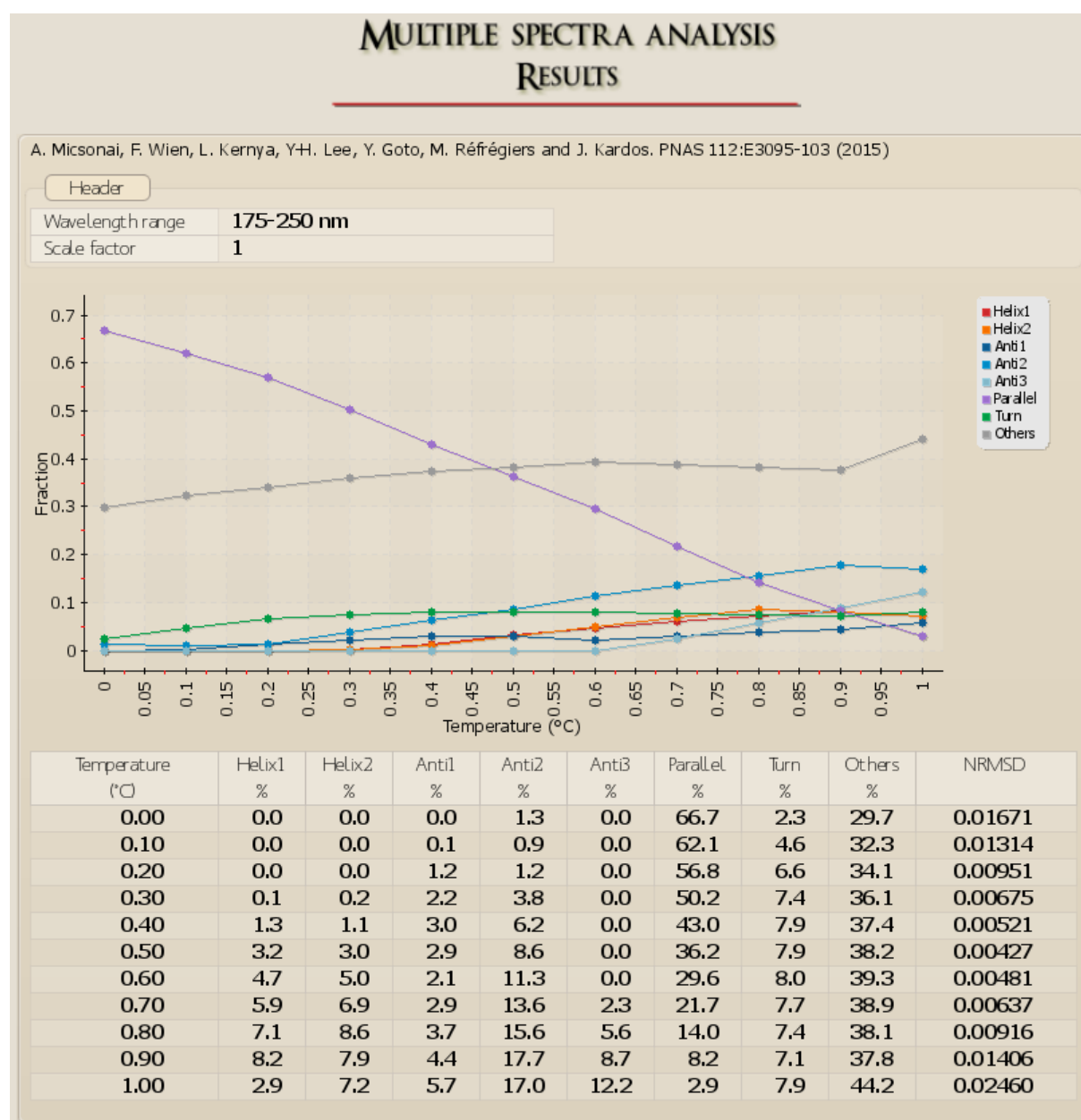
Wavelength (nm)

Experimental.1  
Experimental.2  
Experimental.3  
Experimental.4  
Experimental.5  
Experimental.6  
Experimental.7  
Experimental.8  
Experimental.9  
Experimental.10  
Experimental.11

Calculate the secondary structure

Back

After clicking on the “Calculate the secondary structure” button, the result window will appear.



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.

Choose wavelength range:

☒ 175-250 nm  
☐ 180-250 nm  
☐ 190-250 nm  
☐ 200-250 nm

Scale factor:


Recalculate

Back

## Secondary structure and beta-sheet decomposition for PDB structure module

**SECONDARY STRUCTURE AND BETA-SHEET DECOMPOSITION FOR PDB STRUCTURES**

PDB ID:  
(four letters code, e.g. 1ado)

Please, type the characters above or use your password.

*Secondary structure composition of protein structures deposited in PDB on the basis of the eight structural element of BeStSel. For comparison, DSSP data [Kabsch and Sander, Biopolymers, 22:2577 (1983)] and Selcon3 [Sreerama et al., Protein Sci., 8:370 (1999)] decomposition is also calculated.*

*PDB ID should be given in four letters code format (case-insensitive). At first, results are provided for the entire structure. At the bottom you can select calculations for individual chains. In this case CATH classification [Orengo et al., Structure, 5:1093 (1997)] will also be provided.*

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 element 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. Structures deposited in PDB can be submitted only. PDB ID should be given in four letters code format (case-insensitive).

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

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

**Secondary structure composition**

DSSP <sup>a</sup>	%	BeStSel	%	SELCON <sup>b</sup>	%
3-10-helix	<b>3.3</b>	Helix 1	<b>29.3</b>	Helix 1	<b>29.4</b>
Alpha helix	<b>42.5</b>	Helix 2	<b>13.2</b>	Helix 2	<b>16.4</b>
Pi helix	<b>0.0</b>	Antiparallel 1	<b>0.0</b>	Strand 1	<b>8.3</b>
Turn	<b>10.3</b>	Antiparallel 2	<b>0.0</b>	Strand 2	<b>5.5</b>
Strand	<b>13.8</b>	Antiparallel 3	<b>1.3</b>	Turn	<b>13.8</b>
Bridge	<b>0.6</b>	Parallel	<b>12.5</b>	Others	<b>26.7</b>
Bend	<b>6.5</b>	Turn	<b>10.3</b>		
Loop + missing	<b>23.0</b>	Others	<b>33.4</b>		

**CATH Info<sup>c</sup>**

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

Further references:

<sup>a</sup> : 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

<sup>b</sup> : 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

<sup>c</sup> : 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 first, results are provided for the entire structure in the *Result* page of the „Secondary structure and beta-sheet decomposition for PDB structure” 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., Structure, 5:1093 (1997)] will also be provided (if any).

**Results format**

**Select data to display**

Entire structure  
Chain A  
Chain B  
Chain C  
Chain D

**Secondary Structure:**

☒ DSSP  
☒ BeStSel  
☒ SELCON  
☒ CATH Info

Show! Save image Data in text

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



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

PDB ID: **1ADO** Chain: **A** Length: **363**

### Secondary structure composition

DSSP <sup>a</sup>	%	BeStSel	%	SELCON <sup>b</sup>	%
3-10-helix	<b>3.3</b>	Helix 1	<b>29.5</b>	Helix 1	<b>30.0</b>
Alpha helix	<b>42.7</b>	Helix 2	<b>13.2</b>	Helix 2	<b>16.0</b>
Pi helix	<b>0.0</b>	Antiparallel 1	<b>0.0</b>	Strand 1	<b>8.3</b>
Turn	<b>9.4</b>	Antiparallel 2	<b>0.0</b>	Strand 2	<b>5.5</b>
Strand	<b>13.8</b>	Antiparallel 3	<b>1.3</b>	Turn	<b>13.2</b>
Bridge	<b>0.6</b>	Parallel	<b>12.5</b>	Others	<b>27.0</b>
Bend	<b>6.6</b>	Turn	<b>9.4</b>		
Loop + missing	<b>23.7</b>	Others	<b>34.2</b>		

### CATH Info<sup>c</sup>

Class	<b>3</b>	<b>Alpha Beta</b>
Architecture	<b>20</b>	<b>Alpha-Beta Barrel</b>
Topology	<b>20</b>	<b>TIM Barrel</b>

Further references:

<sup>a</sup> : 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

<sup>b</sup> : 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

<sup>c</sup> : 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

The secondary structure composition of the entire structure or the selected chains displayed separately. The detailed descriptions of the structural elements are described in the original papers, and a brief summary can be found in the "Information" part of the page.




## Fold recognition module

**Secondary structure Content (%)**

Helix1:	<input type="text" value="0"/>
Helix2:	<input type="text" value="0"/>
Anti1:	<input type="text" value="0"/>
Anti2:	<input type="text" value="0"/>
Anti3:	<input type="text" value="0"/>
Parallel:	<input type="text" value="0"/>
Turn:	<input type="text" value="0"/>
Others:	<input type="text" value="0"/>
<b>Total:</b>	0.00

Number of residues:



Please, type the characters above or use your password.

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 sums up to 100.0 % and the chain length is provided. This data may come from previous BeStSel analysis of a CD spectrum (see „Single spectrum analysis and fold recognition” module) or from the analysis of a PDB structure (see „Secondary structure and beta-sheet decomposition for PDB structures” 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) search on single domains with secondary structure composition within the expected error of the CD secondary structure analysis (4) weighted K-nearest neighbors search method. For information on these methods please see page 13-14 in this tutorial, or in the *Information* on the main BeStSel page.

FOLD RECOGNITION RESULTS						
<b>Closest structures in the entire PDB</b>						
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) 2.60.120(20)	2.60.120(30)	6.1.1* (1)	NA	NA
<small>*: 6.1.1. means Unclassified CATH Domain. Numbers in brackets indicate the number of a given CATH domain in the structure.</small>						

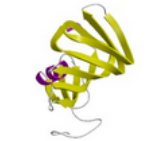


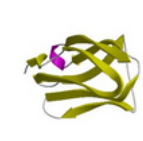


<b>Closest single CATH domains in a PDB subset filtered for &lt;= 90% sequence homology.</b>								
The ten closest structures based on the euclidean distance in the eight dimensional space of BeStSel.								
No.	PDB	Chain	C	A	T	Class	Architecture	Length
1	1VTZ	2	2	60	120	Mainly Beta	Sandwich	196
2	1J58	A	2	60	120	Mainly Beta	Sandwich	384
3	1JHJ	A	2	60	120	Mainly Beta	Sandwich	171
4	1G6E	A	2	60	20	Mainly Beta	Sandwich	87
5	2PHL	A	2	60	120	Mainly Beta	Sandwich	397
6	2W2A	A	2	40	128	Mainly Beta	Beta Barrel	194
7	1LED	A	2	60	120	Mainly Beta	Sandwich	243
8	1ZKO	A	2	40	50	Mainly Beta	Beta Barrel	135
9	1QWR	A	2	60	120	Mainly Beta	Sandwich	319
10	1B35	B	2	60	120	Mainly Beta	Sandwich	255

### Structural domains within the expected error of BeStSel

The algorithm searches on a PDB subset with  $\leq 90\%$  sequential homology for all the chains having single CATH domains that lie within  $1.5 \times \text{RMSD}$  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:

Class			
No.	CATH	Category name	Score
1	2	Mainly Beta	17.34
Architecture			
No.	CATH	Category 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	Category 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	Category 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 4.2 (see the number of domains and categories in the table below). In each layer (Class, Architecture, Topology, Homology) the predicted categories are ordered by the WKNN score calculated excluded every structure belongs to an already predicted categories (lower numbered hits). The WKNN score defined by sum of the weighted distance of every structures (from the query point) among the K- nearest neighbors which belong to the certain category.

Number of	CATH 4.2
Domains	55350
Class	4
Architecture	41
Topology	1310
Homology	5398