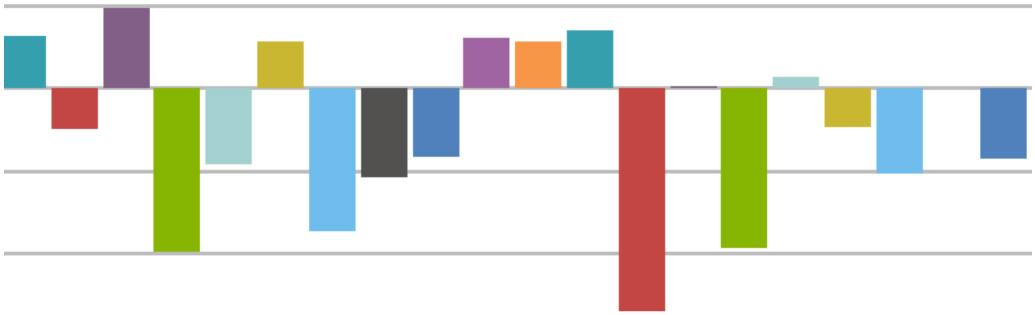
User Guide

Protein|Clpper

Statistical scoring of protease cleavage sites



Content

Version 1.1

November 2015

1. Introduction

Protein|Clpper is a web-based software tool to reveal and analyze amino acid cleavage patterns of proteases from mass-spectrometric data. The analysis is based on biochemical *in vitro* digestion reactions, where substrate proteins are degraded by a protease of interest. The prepared samples containing product peptides are desalted and subjected to tandem mass spectrometry to generate fragmentation spectra that are subsequently analyzed by a proteomics bioinformatics platform (currently Proteome Discoverer, Thermo Scientific, is suited well). For detailed information about experimental settings and an application example, please see the method section in Gersch, M.; Stahl, M.; Poreba, M.; Dahmen, M.; Dziedzic, A.; Drag, M., Sieber; S.A.; "Barrel-shaped ClpP proteases display attenuated cleavage specificities"; *ACS Chemical Biology* **2015**, accepted.

Protein|Clpper is able to display cleavage patterns in two different ways. Firstly, it is connected to the Weblogo service (Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res* **2004**, *14*, 1188), a web-tool to generate multiple sequence alignments of cleavage sites. Conserved preferences are depicted by the height of amino acid letter bars. However, Weblogo does not take into account the differences in the natural occurrences of each amino acid in a respective protein, e.g. leucin is usually an often-used amino acid in many proteins and will therefore be found more frequently in a cleavage site than e.g. tryptophan (assuming a nonspecific cleavage reagent). To correct for this artifact, we developed the Protein|Clpper score S.

The analysis workflow is as follows: First, peptides are discarded if they are present in control runs lacking protease or if they cannot be unambiguously assigned to a protein. Each remaining peptide is interpreted as two cleavage events (or one cleavage event if the peptide is at the protein N- or C-terminus). Next, the number of occurrences of each amino acid **a** at the positions **p** around the cleavage events (P3, P2, P1, P1', P2', P3') is determined both from the amino acids in the peptide sequence itself and from mapping the peptide onto the protein sequence and inferring from there the amino acids of the other side of the cleavage ($\#AA\ per\ position_{\boldsymbol{a,p}}$). These numbers are then divided by the total number of amino acids measured at the respective positions (which is equal to the number of cleavage

events studied for P1 and P1'), yielding the percentage abundance of each amino acid at the respective positions. Protein|Clpper then defines a score $S_{a,p}$ for which these measured abundances are divided by the respective amino acid abundances according to the protein sequence. The measured amino acid frequencies at the respective positions around the cleavage site are thus normalized to the amino acid composition in the protein.

$$S_{a,p} = \frac{\left[\frac{\#AA\ per\ position_{a,p}}{measured\ positions_{p}}\right]}{\left[\frac{\#AA\ per\ protein_{a}}{protein\ length}\right]} = \frac{\%\ of\ amino\ acid\ \boldsymbol{a}\ at\ position\ \boldsymbol{p}\ measured}{\%\ of\ amino\ acid\ \boldsymbol{a}\ expected\ from\ protein\ sequence}$$

The interpretation of this score is as follows:

$\log_2(S) > 0$: amino acid is enriched at position

$\log_2(S) \approx 0$: amino acid found as expected from random cleavage

$\log_2(S) < 0$: amino acid is depleted at position

In order to integrate data from different runs and different substrate proteins, Protein|Clpper averages the score values of all measured proteins and weights them according to the measured peptide spectral matches per protein. Proteins with less than 10 cleavage events are excluded.

Sequence coverages, Weblogos and scores are interactively presented on the Protein|Clpper website. Additionally, all data (different types of scores and Weblogo consensus sequences) can be downloaded as csv files. Data will not be stored permanently on the server.

2. Protein|Clpper Analysis Procedure

To use Protein|Clpper, please conduct the respective experiments described in the section above and analyze your data with Proteome Discoverer (Thermo Scientific). Protein|Clpper has already been tested with proteome Discoverer version 1.4. Then export your results to a txt file including all PSMs.

(1) Upload txt exports and FASTA database to the Protein|Clpper environment.
There are two types of txt data: A) Your original runs and B) optionally also control runs (e.g. where a catalytically inactive variant of the protease was used). More than

one file at the same time may be uploaded and concatenated by the system. Additionally, it is necessary to add a FASTA database (.fasta file) that contains the sequences of all proteins present in the sample, so that,the software is able to map peptides to the complete protein sequences.

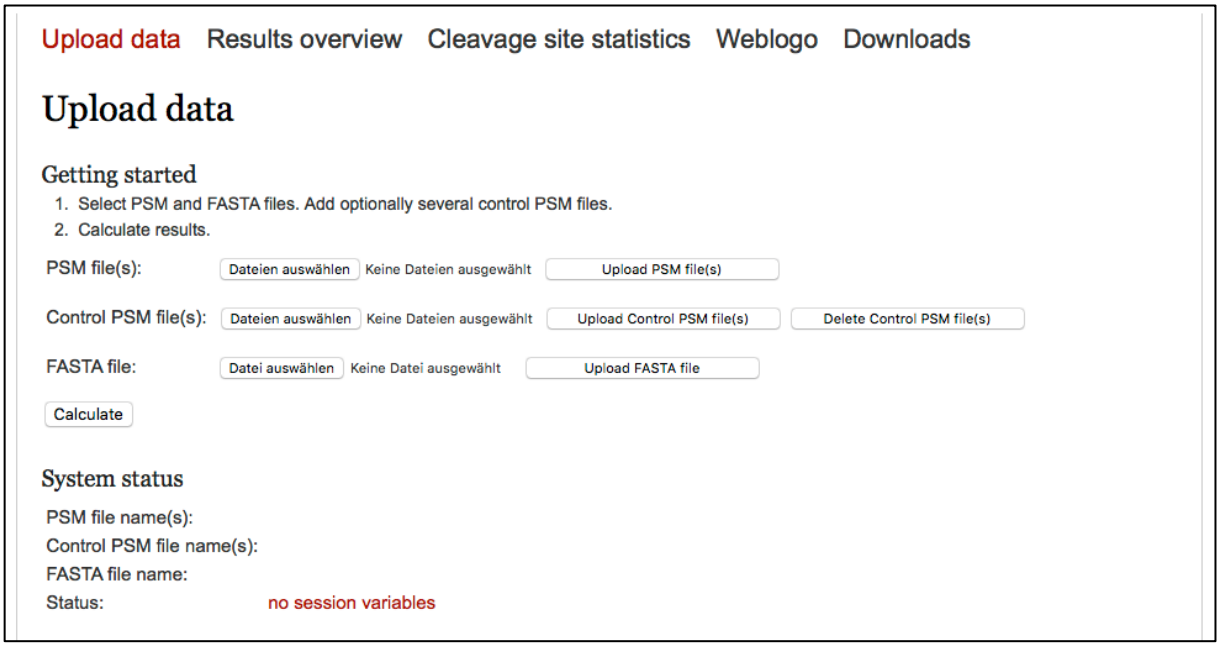


Figure 1: "Upload data" section. Click on "select files" (label may change depending on your country) and upload them by pressing the "upload" buttons (pressing one of the buttons is sufficient to upload all files).

After uploading all files, the system status will change.

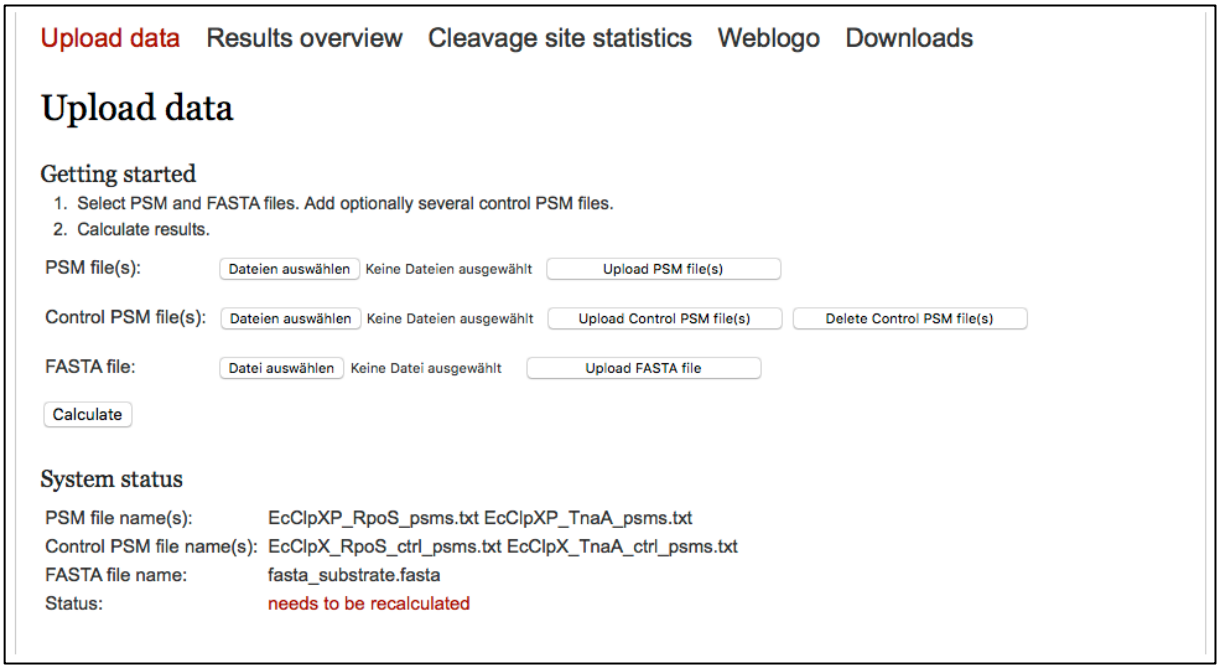


Figure 2: Change of the system status.

(2) Calculation of cleavage properties.

Click the “Calculate” button and the system status will indicate successful calculations. However, if calculations are not possible, please check that your files contain the required columns (see section 3).

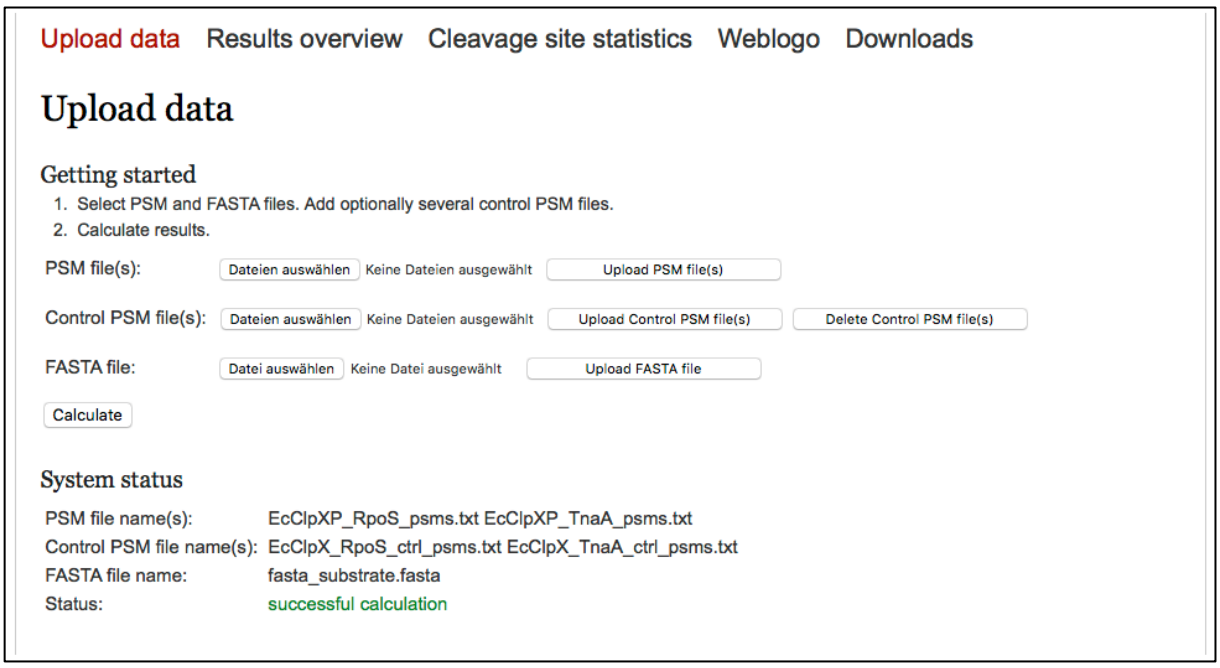


Figure 3: All calculations were conducted successfully.

(3) General results overview.

Click on “Results overview” to display general information about your data.

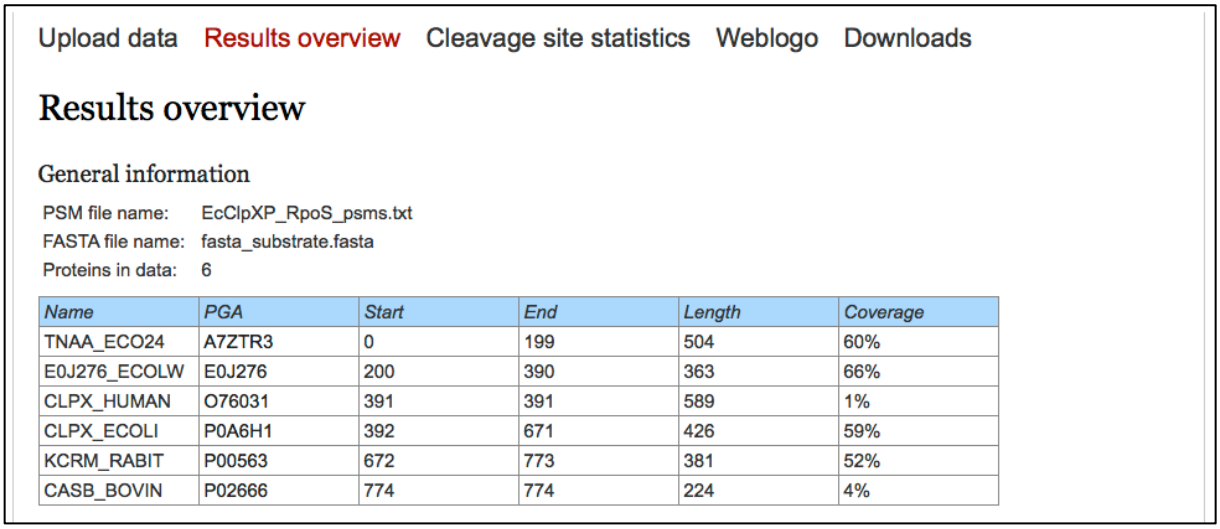


Figure 4: General results are presented in the “Results overview“ section.

All found proteins are listed under “General information”. “PGA” represents the protein accession code for the Uniprot database, if you have used protein data from there. Other databases may be used as well. “Start” and “End” indicate internal

counting IDs for respective PSMs (peptide spectral matches). “Length” indicates the total number of amino acids in this protein and “Coverage” displays the sequence coverage produced by the input peptides subtracted by control peptides.

Information to single proteins

normal weighted unique peptides

TNAA_ECO24	*Cleavages*	*Peptides*	*Weights*
	400	200	928.32

E0J276_ECOLW	*Cleavages*	*Peptides*	*Weights*
	382	191	1295.68

CLPX_HUMAN	*Cleavages*	*Peptides*	*Weights*
	2	1	3.28

CLPX_ECOLI	*Cleavages*	*Peptides*	*Weights*
	560	280	1490.38

KCRM_RABIT	*Cleavages*	*Peptides*	*Weights*
	197	102	510.12

CASB_BOVIN	*Cleavages*	*Peptides*	*Weights*
	2	1	3.24

Averaged	*Cleavages*	*Peptides*
	1539	773

Figure 5: Information about single proteins.

“Information about single proteins” is a list of all detected proteins and the respective number of cleavages and peptides. “Weights” indicates the sum of all XCorr values. “Unique peptides” are refered to one PSM per distinct peptide.

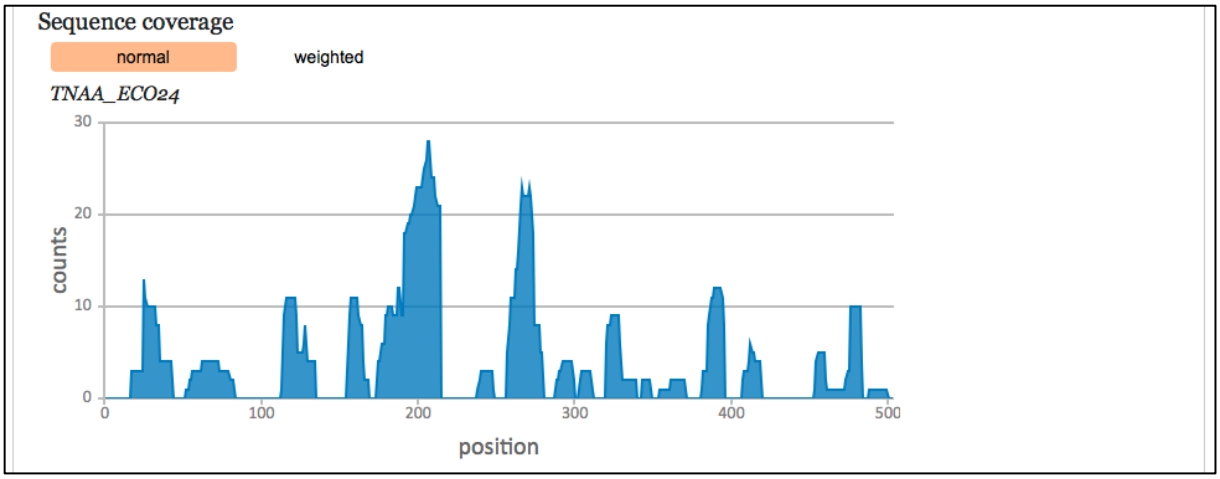


Figure 6: Coverage plot.

Furthermore, the sequence coverage is plotted against the whole protein sequence resulting in a coverage plot. This gives an overview of the peptide distribution on a specific protein.

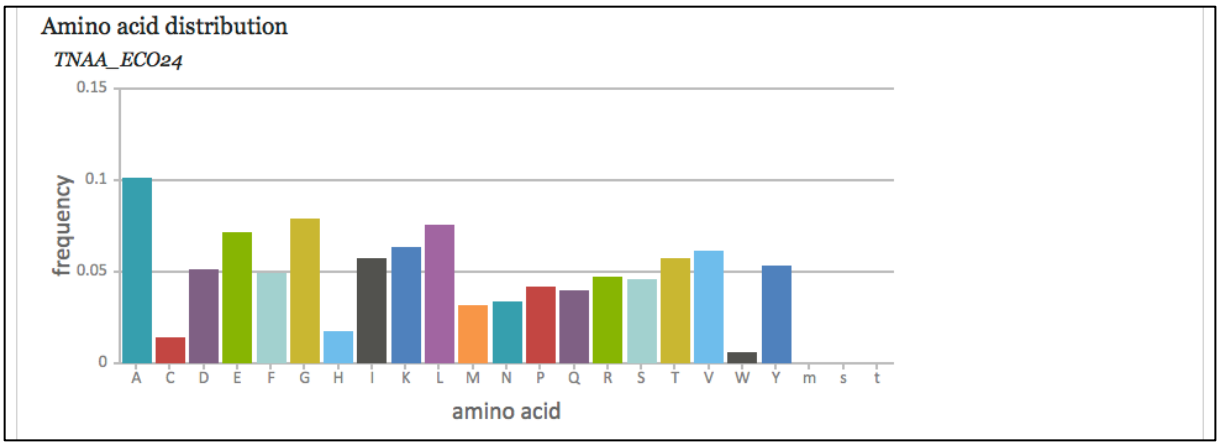


Figure 7: Amino acid distribution.

The amino acid distribution plot shows the portion of each amino acid in the full protein amino acid sequence. (m, s, t are amino acid variants that are not used, yet)

(4) Detailed view of the cleavage sites.

Click in the main navigation on “Cleavage site statistics”.

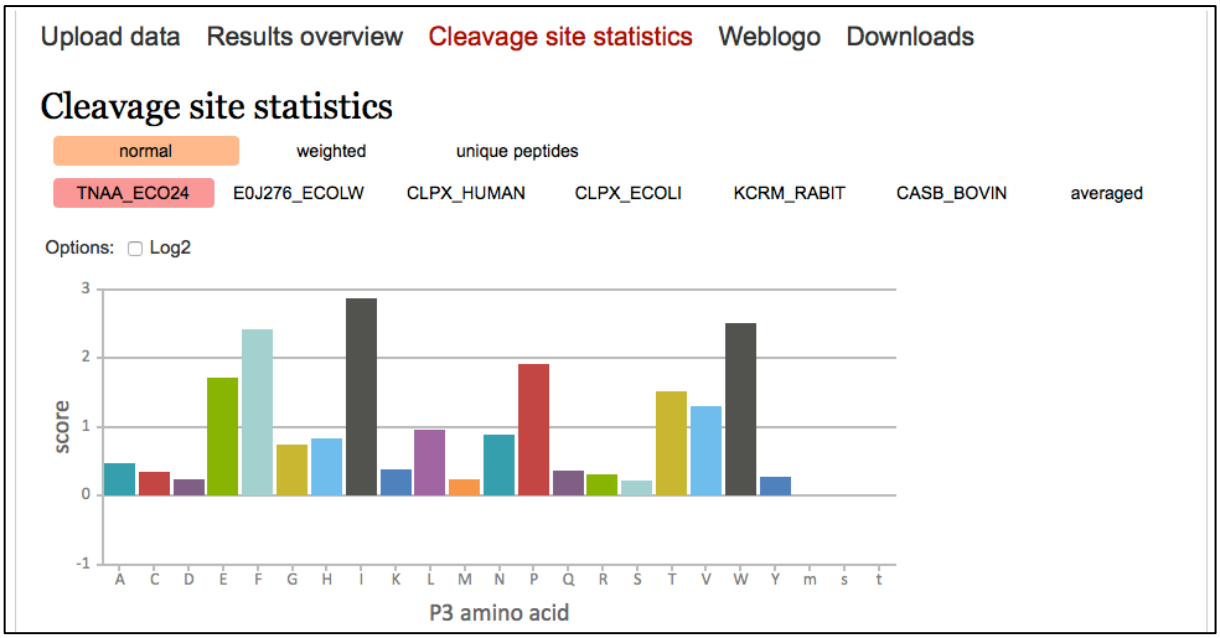


Figure 8: Cleavage site statistics.

First, the cleavage sites can be viewed in “normal” mode. In this case each PSM is encountered. Using the “weighted” mode, the XCorr value is also taken into account. “unique peptides” pools every set of PSMs to one peptide to one PSM. Next, the cleavage sites for each of the found proteins can be viewed by clicking on its ID. Alternatively, an average of all proteins can be displayed. Please note that proteins that were cleaved less than ten times are not included in the average calculation. By checking the “Log2” box, the scores are transformed logarithmically.

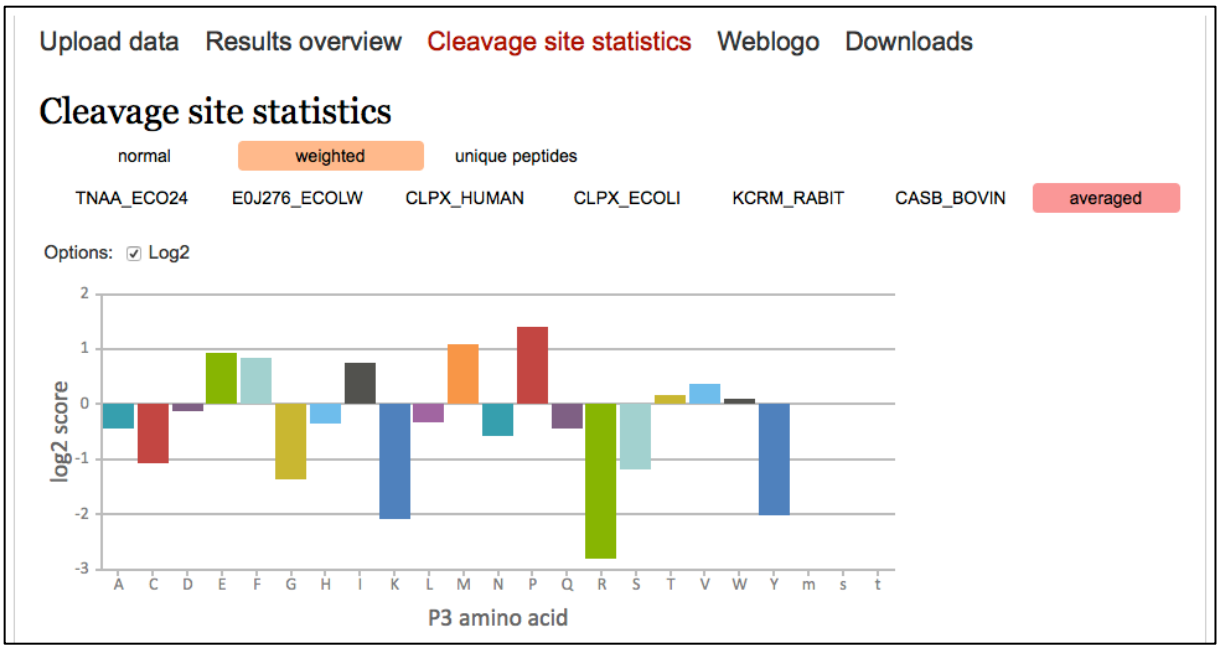


Figure 9: Logarithmic representation of weighted and averaged scores.

Six positions (P3 to P3') around each cleavage site are analyzed.

(5) Generation of Weblogos.

Click on "Weblogo" in the main menu and set the y-axis unit as well as the y-axis scale. Then click on "Generate Weblogo from session data". Subsequently, the averaged data set will be submitted to a Weblogo server and a png will be displayed.

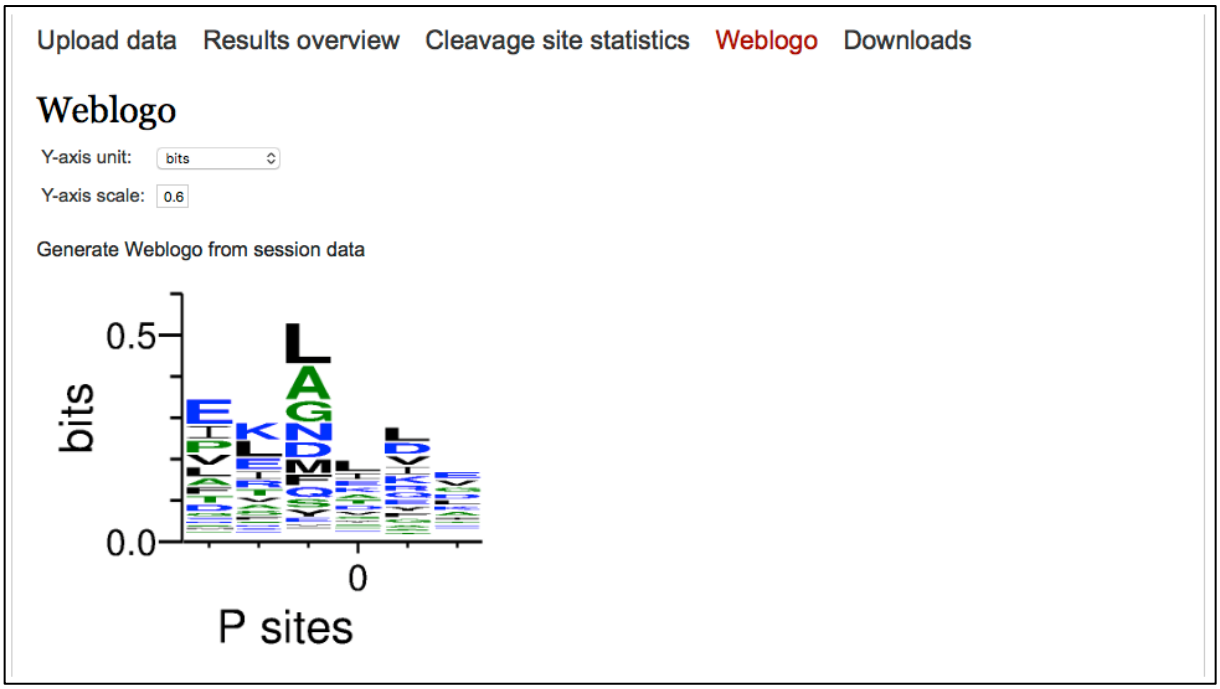


Figure 10: Weblogo of averaged and non-weighted data.

Y-axis units and scale can be changed and the Weblogo will be recalculated by clicking again on “Generate Weblogo from session data”. For more options on Weblogos please use the consensus list download described below.

(6) Download of the complete data set.

By clicking on “Downloads”, there is the possibility to access calculation data by spreadsheet programs like Microsoft Excel.

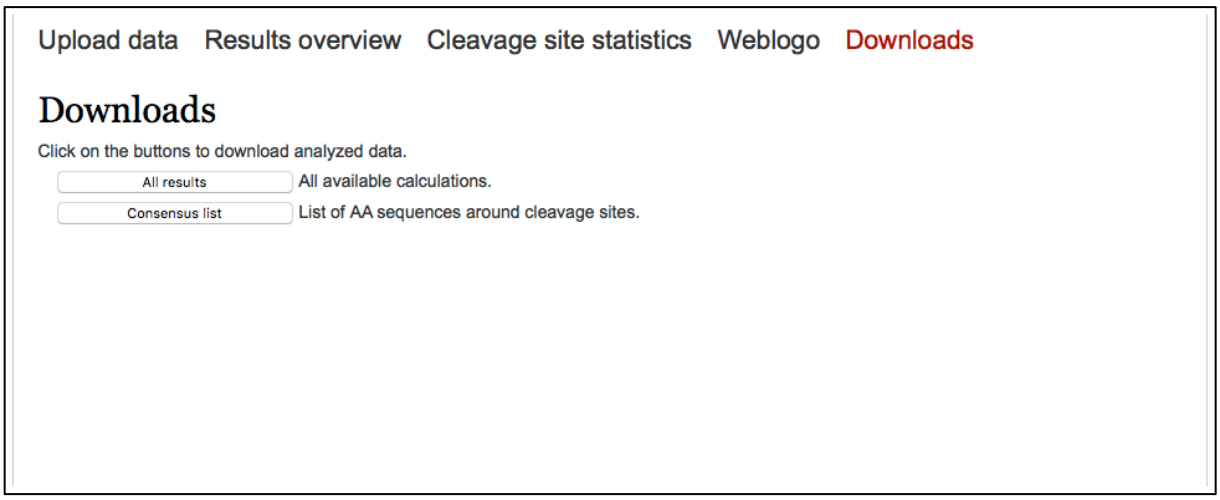


Figure 11: Download interface.

The files are in csv format and columns are separated by commas.

3. Input and Output Files

Input.

Txt files containing peptide spectral matches: At least the following columns should be included: Confidence level, Sequence, PSM Ambiguity, # Protein Groups, Protein Group Accessions, XCorr. The FASTA should be designed according to the file type specifications. FASTA headers should look like this:

```
>sp|P02662|CASA1_BOVIN Alpha-S1-casein OS=Bos taurus GN=CSN1S1 PE=1 SV=2
```

Output.

A complete data set is provided in the xxx__all_results.csv. The table is subdivided in six cleavage site positions P3 to P3’. Basically, data is a numerical representation of the graphs of the section “Cleavage site statistics”. The xxx__consensus_list.csv contains a list of all cleavage sites as six amino acid windows. This list can directly

be provided as input on the Weblogo webpage (http://weblogo.threeplusone.com/create.cgi).

4. Contact Information

Protein|Clpper was developed at

Technische Universität München,
Department of Chemistry,
Lichtenbergstr. 4,
D-85747 Garching, Germany.
www.oc2.ch.tum.de

In case of questions, comments and requests, feel free to contact matthias.stahl@mytum.de, malte.gersch@mytum.de or stephan.sieber@mytum.de

If you find Protein|Clpper useful for your research, please cite:

Gersch, M.; Stahl, M.; Poreba, M.; Dahmen, M.; Dziedzic, A.; Drag, M., Sieber; S.A.; “Barrel-shaped ClpP proteases display attenuated cleavage specificities”; *ACS Chemical Biology* **2015**.