CleaveScope (v. 1.0)

User Manual

CleaveScope is a Java-based software designed for protease specificity analysis using LC–MS/MS data. It allows users to process Mascot XML files, visualize amino acid frequency matrices as heatmaps, and generate sequence logos based on Z-score or normalized values.

1. Installation

CleaveScope is distributed as a standalone executable file (CleaveScope.exe) that does not require Java to be preinstalled on the system. The program includes its own embedded runtime environment, so it can be launched directly on any Windows computer. Launch the application by double-clicking the executable. If Windows SmartScreen warns about an unknown publisher, select "Run anyway" (the application is safe and unsigned, source code available at https://github.com/MiKonstantinov/CleaveScope).

System requirements: Windows 7 or later (64-bit), at least 4 GB RAM. No Java installation needed.

Alternatively, CleaveScope can also be launched from the command line using a JAR file (file cleavescope-1.0-SNAPSHOT-jar-with-dependencies.jar in folder runtime) if JavaFX SDK is installed:

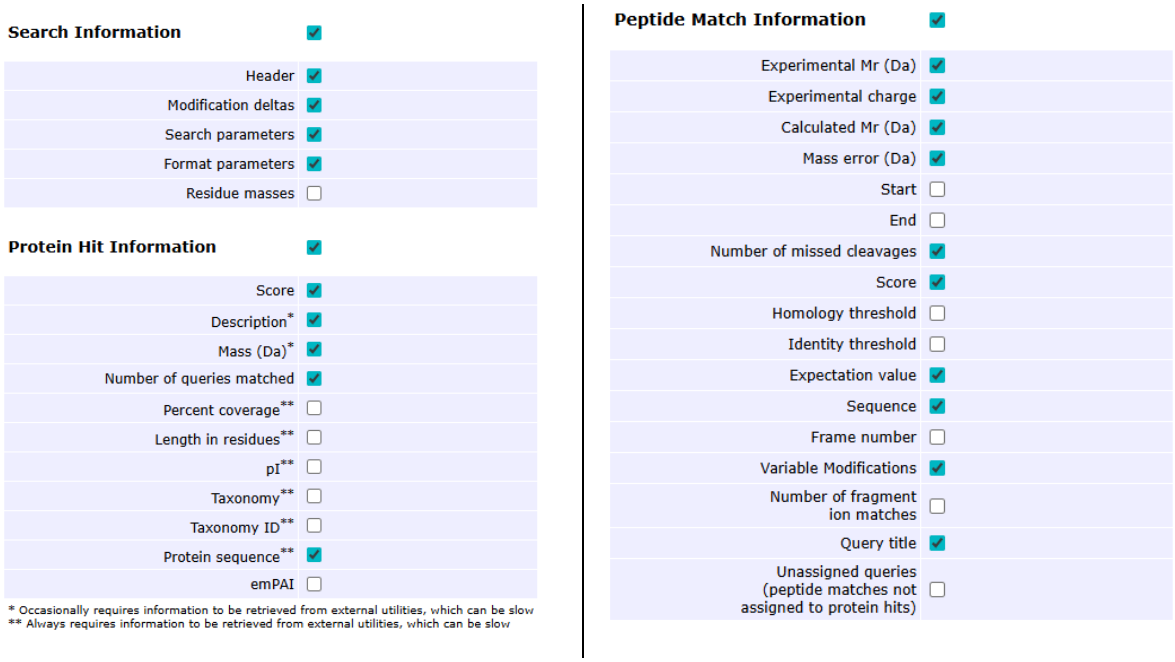
```
java --module-path "C:\Program Files\Java\javafx-sdk-23.0.1\lib" --add-modules javafx.controls,javafx.fxml,javafx.web -jar cleavescope-1.0-SNAPSHOT-jar-with-dependencies.jar
```

To use this method, you need to have Java 17 or later and JavaFX SDK 23.0.1 (or newer) installed, with the correct path to the JavaFX library specified in the --module-path argument.

2. Input Data

CleaveScope processes multiple Mascot search result files in XML format. To ensure compatibility, the exported XML files must include protein sequences (disabled by default in Mascot) and peptide sequences.

Export the data using mostly default Mascot settings, as shown in the example below. Make sure the following options are enabled in the export window.



Important! Because CleaveScope determines protease specificity independently, the Mascot search must be performed without specifying an enzyme (Enzyme: None).

3. Data Analysis

After exporting the XML files, open the Data Analysis tab in CleaveScope. Here, you can:

- Select one or more XML files for analysis;
- Define the size of the positional window around the cleavage site according to the Schechter and Berger nomenclature (e.g., from P_6 to P_5' by default, or up to P_{12}–P_{12}', where P_1 represents cleavage site).
- Optionally exclude N-terminal methionine (N-Met) from analysis, since the removal of the initiator methionine represents a natural post-translational modification.

CleaveScope extracts all unique peptides from the imported Mascot XML files, removing duplicates. For each peptide, the program identifies its cleavage site position within the parent protein by matching the peptide sequence to its corresponding region in the full protein sequence. Using this information, CleaveScope reconstructs the amino acid environment surrounding cleavage site (P_1) and aligns all residues relative to this position. For every identified site, amino acid residues are extracted within the user-defined positional window. If a peptide is located near the protein terminus and lacks sufficient upstream or downstream residues, the missing positions are automatically filled with the placeholder symbol “Z” to preserve alignment length.

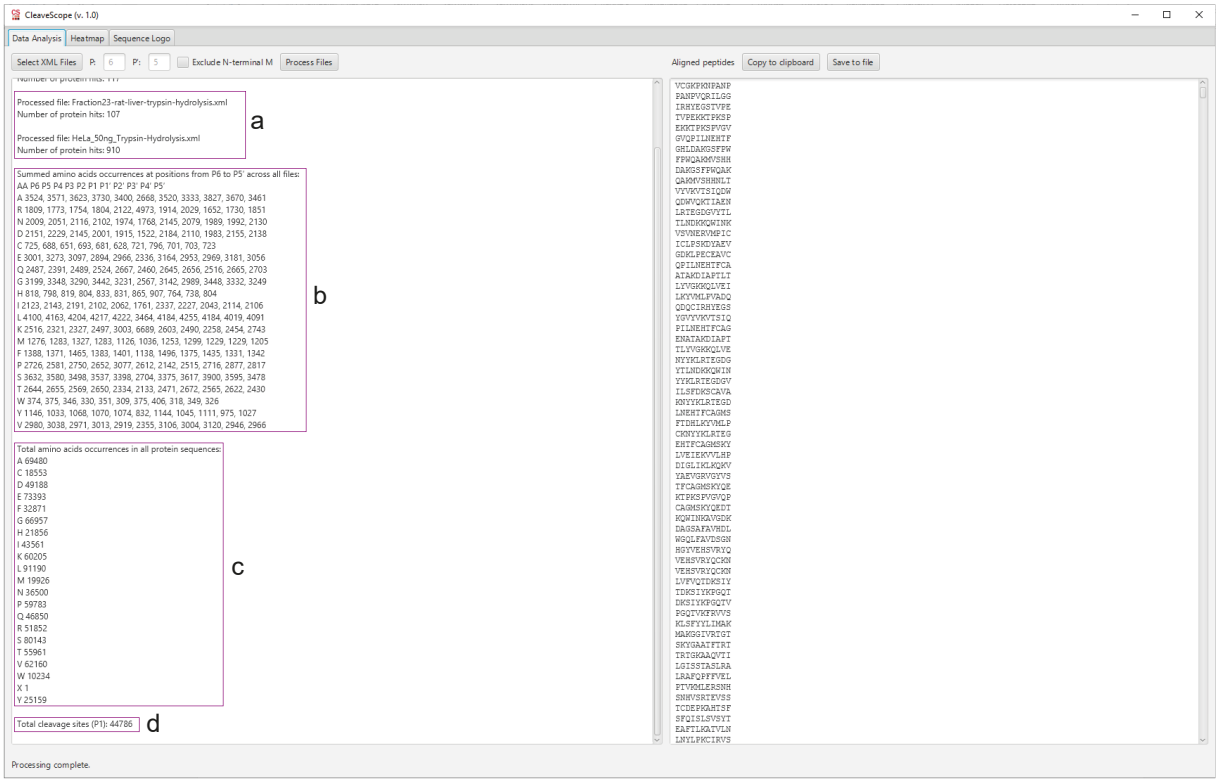
Data processing begins when the “Process files” button is pressed in the Data Analysis tab. After processing, the left panel displays a summary of the analyzed data, including:

a. the name of each processed file and the number of protein hits;

b. the matrix of amino acid occurrences at each selected position (e.g., from P6 to P5′);

c. the overall amino acid distribution calculated from all proteins across all files;

d. the total number of unique cleavage sites identified in the dataset.



The resulting dataset – including both the positional frequency matrix and the background composition – is then used for visualization in the Heatmap and Sequence Logo tabs, where protease specificity can be analyzed and compared under different normalization modes.

4. Visualization

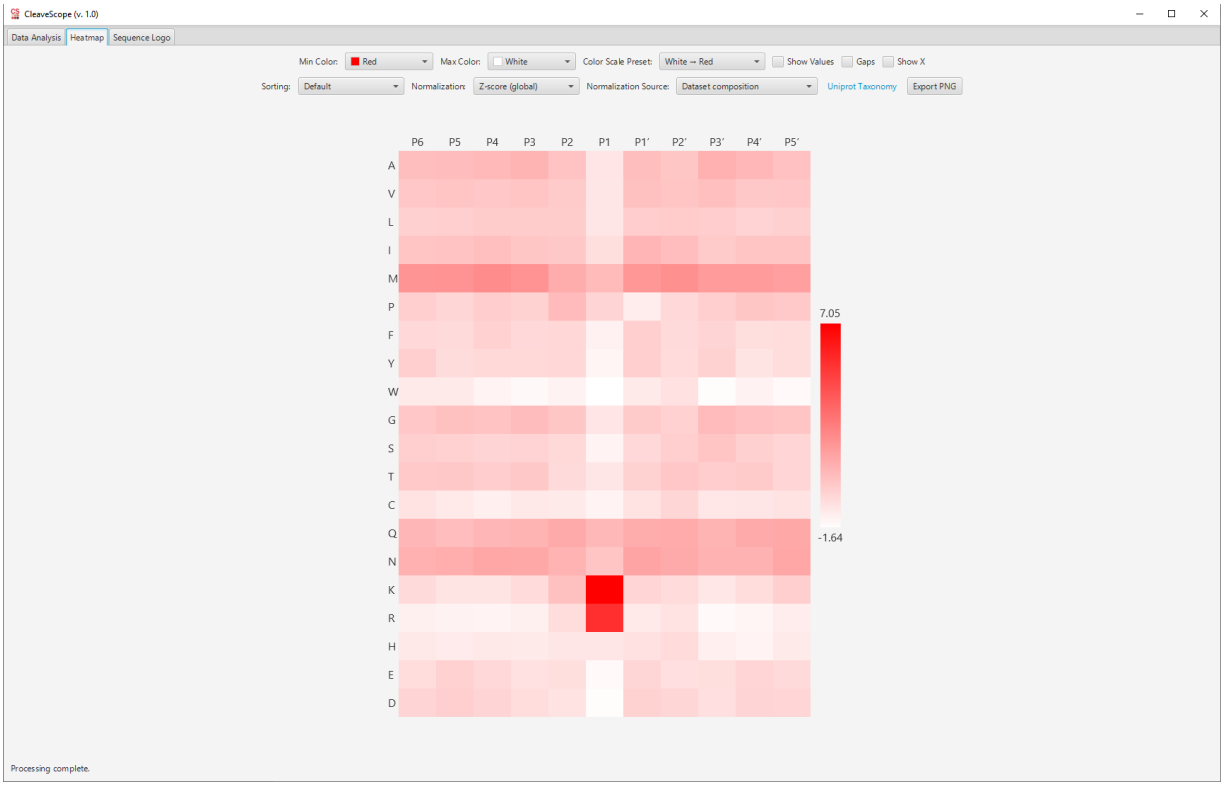
CleaveScope provides two main visualization modes: Heatmap and Sequence Logo. Both are accessible through the upper tab bar and use the processed dataset (positional frequency matrix and background composition) generated during data analysis.

The Heatmap tab displays a matrix of amino acid frequencies at each position relative to the cleavage site. Each cell corresponds to a value describing how frequently a given amino acid occurs at a specific position, and the intensity of color reflects this frequency. The color scale ranges from the minimum to the maximum value, displayed on the right side of the map.

At the top of the heatmap window, several controls allow the user to adjust visualization parameters:

- Min/Max Color: manual selection of the gradient endpoints;
- Color presets: predefined gradients such as White → Red etc.;

- Show Values: displays numeric values inside cells;
- Gaps: includes or excludes gaps between cells;
- Show X: toggles visibility of the undefined residue “X” if present;
- Sorting: controls the amino acid order in rows – Default, Alphabetical, Hydrophobicity or Polarity;
- Normalization: choice of method used to scale the data – Z-score (global), Z-score (row), Direct normalization, Percentage (per residue) or Raw values;
- Normalization Source: defines whether normalization is based on the current dataset composition or on an external background – either loaded from a FASTA file, entered as a custom amino acid composition, or retrieved automatically from Swiss-Prot (for *Homo sapiens*, *Mus musculus*, *Escherichia coli*, or a user-defined NCBI Taxon ID).



CleaveScope supports several normalization approaches to compare amino acid preferences under different analytical conditions. Each method transforms raw amino acid counts into values suitable for visual interpretation and statistical comparison.

Direct normalization – calculates the ratio of each amino acid frequency at position j to its overall background frequency (bg_i), either from the analyzed dataset or from an external FASTA/Swiss-Prot database:

$$f_{ij}^{\text{norm}} = \frac{f_{ij}}{bg_i}$$

where f_{ij}^{norm} is the normalized value, f_{ij} is the absolute number of observations of amino acid *i* at position *j*, and bg_i is the total number of amino acid *i* occurrences in the background dataset.

This method highlights residues that occur more or less frequently than expected based on their general abundance.

Z-score (global) – standardizes the directly normalized matrix across all positions and amino acids:

$$Z_{ij}^{global} = \frac{f_{ij}^{norm} - \mu}{\sigma}$$

where Z_{ij}^{global} is the Z-score of amino acid *i* at position *j*, μ is the mean of f_{ij}^{norm} across all amino acids and all positions, σ is the standard deviation of f_{ij}^{norm}across all amino acids and positions, and f_{ij}^{norm} is the normalized value obtained from direct normalization.

It emphasizes statistically over- or underrepresented residues and facilitates comparison between positions.

Z-score (row) – normalizes values within each amino acid separately, showing in which positions a given residue occurs more or less frequently compared to its own average occurrence across all positions.

$$Z_{ij}^{row} = \frac{f_{ij} - \mu_i}{\sigma_i}$$

where Z_{ij}^{row} is the Z-score of amino acid *i* at position *j*, μ_i is the mean of f_{ij}^{norm} across all positions for amino acid *i*, σ_i is the standard deviation of f_{ij}^{norm} across those positions, and f_{ij} is the absolute number of observations of amino acid *i* at position *j*.

It shows in which positions a given residue occurs more or less frequently than the average for the same amino acid in other positions.

Percentage (per residue) – expresses how each amino acid is distributed across all positions:

$$P_{ij} = \frac{f_{ij}^{norm}}{\sum_k f_{ik}^{norm}} \times 100\%$$

where P_{ij} is the percentage of amino acid *i* at position *j*, f_{ij}^{norm} is the normalized value obtained from direct normalization, and $\sum_k f_{ik}^{norm}$ is the sum of normalized values of amino acid *i* across all positions.

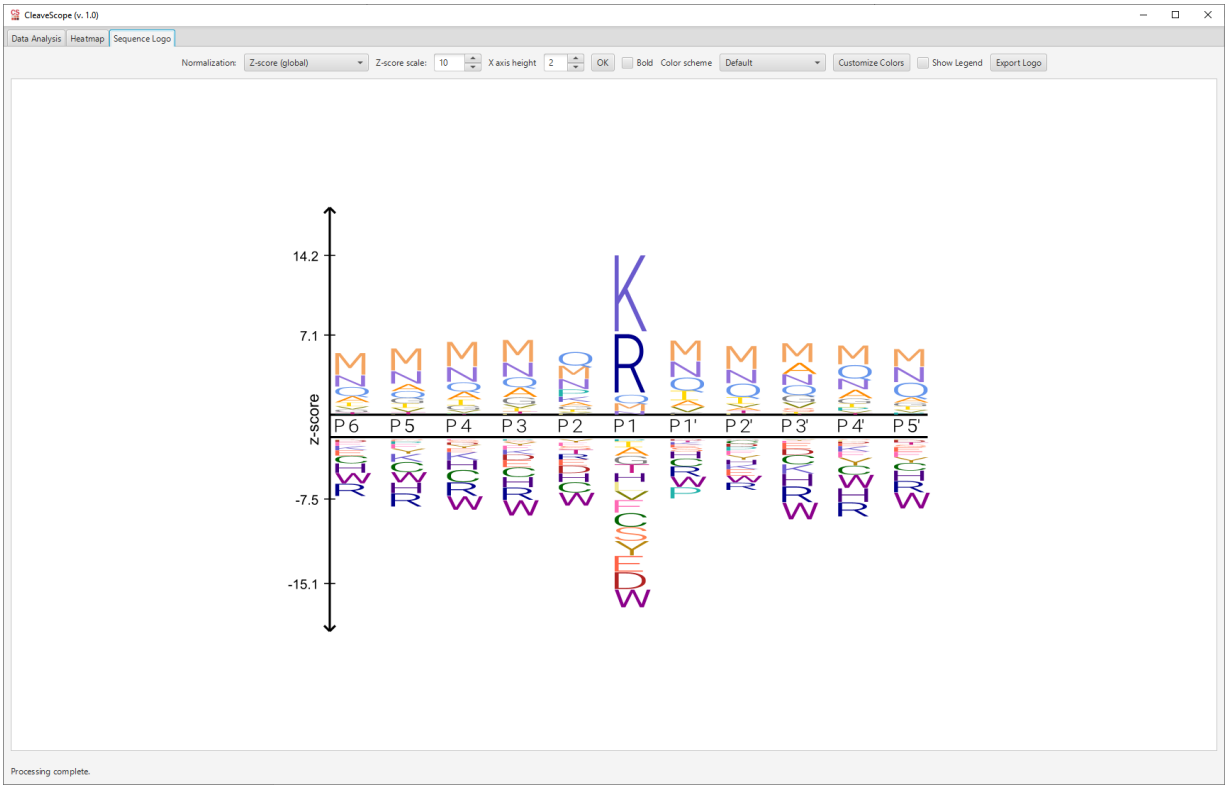
This mode allows comparing positional trends within the same amino acid type, independent of overall abundance.

These normalization methods can be combined with different background sources (dataset, FASTA, or Swiss-Prot), providing flexibility in visualizing protease specificity and comparing experimental data with known proteomes.

The Sequence Logo tab provides a complementary visualization of protease specificity based on the same processed data matrix used for the heatmap. Each position around the cleavage site is represented by a stack of amino acid symbols whose heights correspond to their relative occurrence or statistical weight, depending on the selected normalization mode.

At the top of the Sequence Logo tab, a control panel allows users to adjust the appearance, scaling, and coloring of the logo:

- Normalization: selects which matrix is used to generate the logo – Z-score (global) or Z-score (row). It is not applicable to other normalization methods or to Raw values, since those representations do not provide statistically comparable scales for letter height calculation;
- Z-score scale: defines the vertical scaling factor (higher values increase the relative height of letters);
- X axis height: sets the overall height of the X-axis;
- Bold: toggles bold font rendering for amino acid letters;
- Color scheme: controls how amino acids are colored in the logo. Available presets include: Default, Group-based, Hydrophobicity, Grayscale and Random;
- Customize Colors: opens a dialog for manually changing color assignments for each amino acid;
- Show Legend: toggles the display of the color legend beneath the logo.



5. Export

CleaveScope allows exporting both numerical data and graphical visualizations for further analysis or publication use.

Program supports exporting aligned fragments in text format (.txt) from the Data Analysis tab.

From the Heatmap tab, users can export the current visualization as a PNG image using the Export PNG button. The exported image includes all labels, positional axes (P and P′), and the corresponding color scale, preserving the current normalization, gradient, and sorting settings. In the Sequence Logo tab, the logo can be saved as either SVG or PNG using the Export Logo button. The SVG format preserves vector quality, color coding, transparency, and scaling, making it suitable for use in figures and publications.

Important! The exported image size for both the heatmap and the sequence logo depends on the current window size. For best quality, it is recommended to switch the program to fullscreen mode before saving the images.