OverProt Server manual

OverProt Server (OVERview of PROTein family) is a web application generating an overview (i.e., a consensus sequence) of protein family secondary structure elements (SSEs). The application consists of four pages – main introductory one (Main page), page for specification of user-defined queries (Queries page), visualization of the results (Results page), and integrated view of a specific domain (Domain view page). This manual also provides information about OverProt limitations and about the desktop version of OverProt.

Contents

Main page

The main page offers a possibility to **select a protein family** using its CATH ID:

CATH ID is in a format *n.n.n.n*, where *n* denotes a number (e.g., 2.20.20.10 for Anthopleurin-A protein family). Clicking on "Go" button will retrieve the precomputed consensus SSE sequence and redirect to the Results page. The search box also allows searching for a specific CATH domain (e.g., 2nnjA00) or PDB entry (e.g., 2nnj).

The search box also provides suggestion functionality, showing the IDs and names of existing CATH families, organized by CATH Class, Architecture and Topology (the number of suggestions is limited to 10), e.g.:

The main page also enables the user to decide for selection of protein family members included into computation of consensus sequence. It can be done via **user-defined queries**:
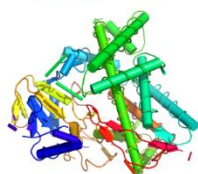
User-defined queries

Run OverProt on your own list of protein structures ⇒

By clicking on ⇒, the user is redirected to Queries page.
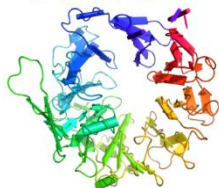
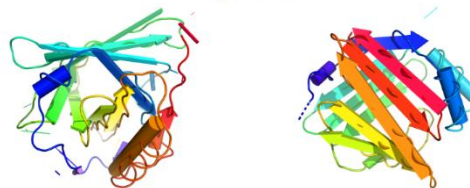Last but not least, the Main page includes **examples**:

Examples

By clicking on the figures of the protein family representatives, the user is redirected to Results page, containing precalculated results for particular protein families.

Queries page

This page includes a form for definition of protein family members included into the calculation of a consensus sequence:

User-defined query

Job name:

Protein domain list:

List domains one-per-line,
PDB,CHAIN or **PDB,CHAIN,RANGES**
(maximum 500 domains) ⓘ

Example:
```
1og2,A
1og2,B
1bu7,A,100:450
1bu7,B,100:178,185:370,390:
```

[Load example]

Expected computation time

or load domain list from file [Choose file]

[Submit]

The user can make a selection of input proteins in two ways – using the **input form** or by **submitting a text file**. These ways cannot be combined.

Selection of input proteins using the **input form**:
The user can provide the following information about an input protein:
- PDB ID of a protein and ID of a chain, for example:
    ```
    1og2,A
    1bu7,B
    ```
- PDB ID of a protein, ID of a chain, and a selection of residues. The selection consists of one or more residue ranges, each range given by the first and the last residue number. The first and/or last residue of a range can be omitted. For example:
    ```
    1bu7,A,100:
    1bu7,A,100:450
    1bu7,B,100:178,185:370,390:
    ```

The chain IDs and residue numbers must be provided following the label_* numbering scheme (mmcif-style, corresponding to columns label_asym_id and label_seq_id in mmCIF files), not the auth_* numbering scheme (PDB-style, corresponding to columns auth_asym_id and auth_seq_id in mmCIF files and columns in traditional PDB files). However, in most cases these numberings are identical.
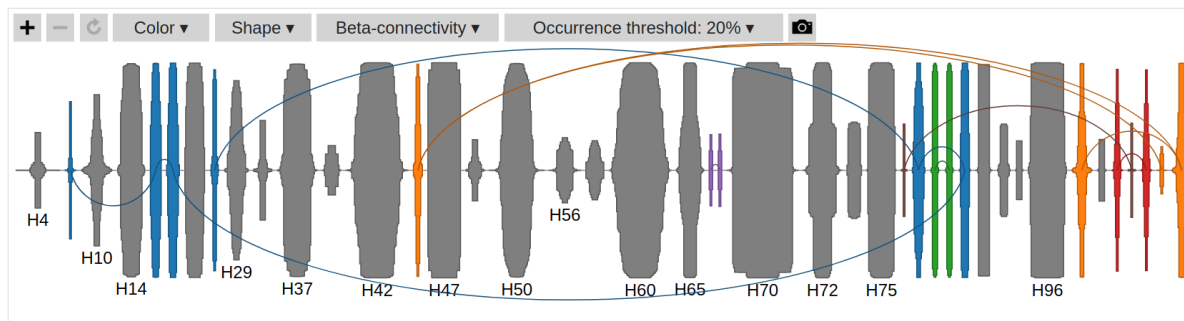
Selection of input proteins by **submitting a text file**:
The same information which the user provides via the input form can be provided also in a text file.

Results page

The Results page contains an interactive visualization of a SSE consensus sequence, for example:

A consensus sequence is depicted as the diagram, where:

- The X-axis shows the order of SSEs, which are depicted as pure rectangles or ovals. The Y-axis indicates the occurrence. As a result, height shows the percentage of protein structures that contain this SSE, while the width indicates the average length measured as the number of amino acid residues. The ovals (Shape:SymCDF visualization style) also represent the variability of the length of the specific SSE – if the length is uniform, the shape is more rectangle-like; if the length is more variable then the shape is deformed according to the statistical distribution of the SSE length.
- Only consensus SSEs with occurrence above the occurrence threshold are shown (default: 20%).
- Rectangles depicting β-strands from the same β-sheet have the same color, while all helices are gray.

Parts of the sequence can **be zoomed in and out** by ➕ ➖ ↺. Hovering over an SSE shows SSE details:
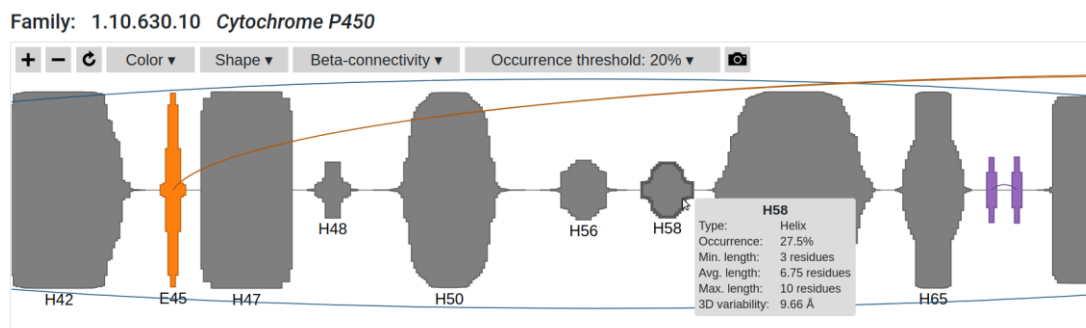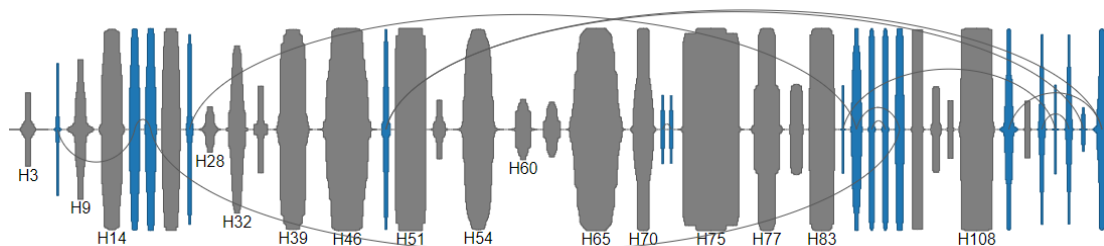

The consensus sequence can be **colored** in several ways using Color ▾, the possibilities are:

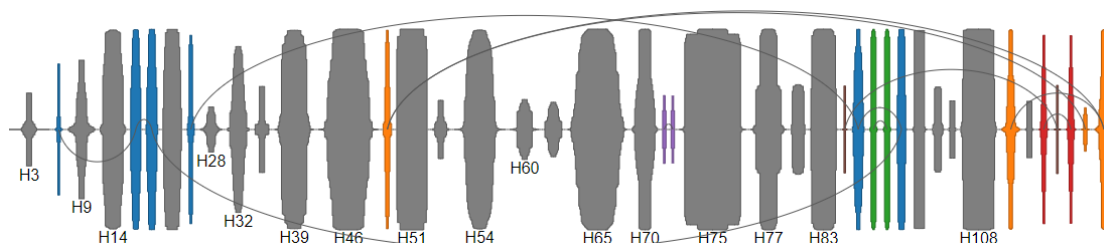- **Uniform:** All SSEs in the same color.
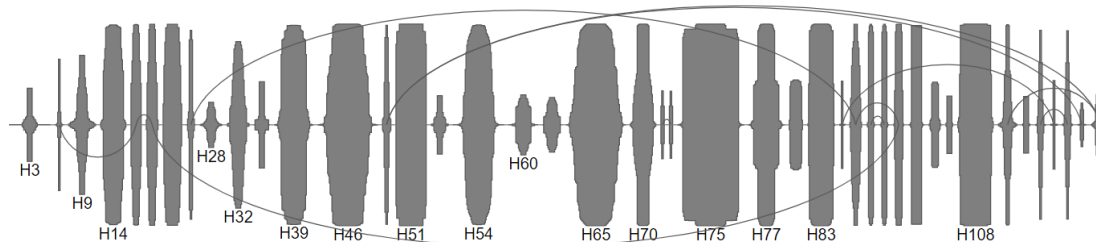
- **Type:** β-sheets in blue, helices in gray.

- **Sheet: Default coloring.** β-strands from the same β-sheet have the same color, while all helices are gray.


- **Variability:** 3D variability measures the standard deviation of the SSE end point coordinates. Low values (dark) indicate conserved SSE position, high values (bright) indicate variable SSE position.
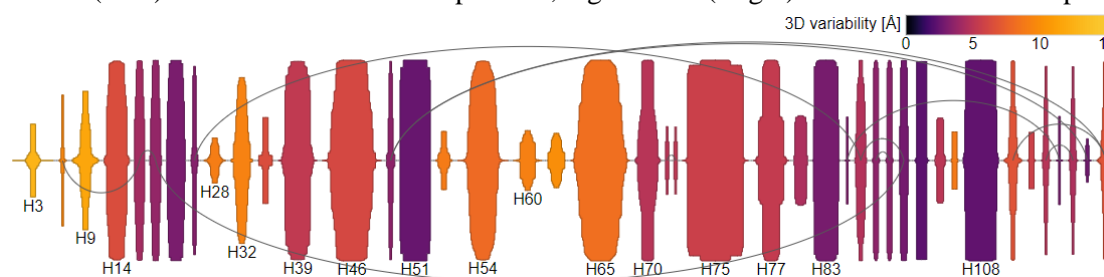

The SSEs can have two **shapes** using � Shape ▾ , the possibilities are:

- **Rectangle:** Show SSEs as rectangles. Height of the rectangle indicates occurrence (what percentage of structures contain this SSE), width indicates average length (number of residues).
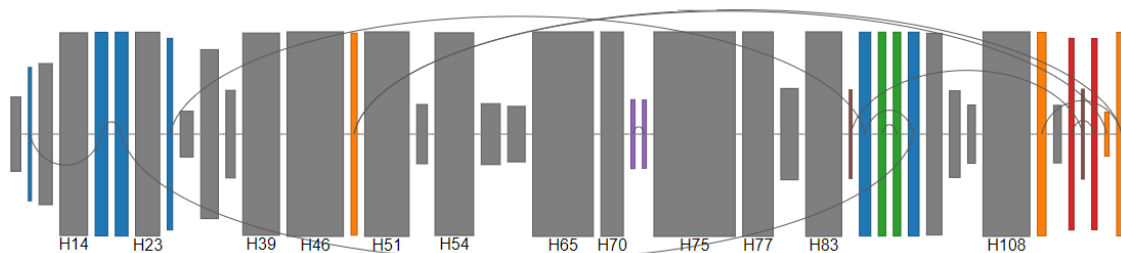

- **SymCDF: Default shape.** Show SSEs as symmetric cumulative distribution function shape. Height of the rectangle indicates occurrence, width indicates maximum length, shape shows length distribution.
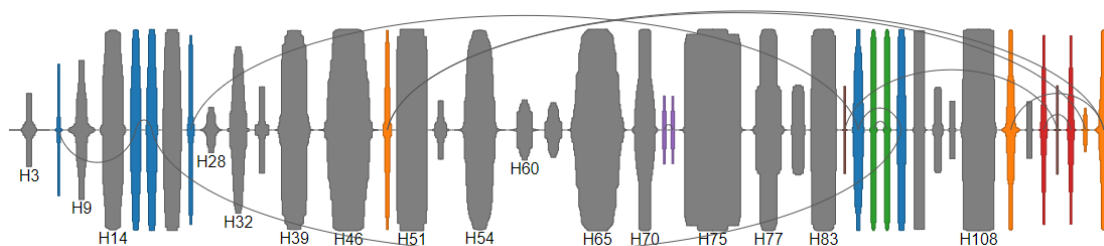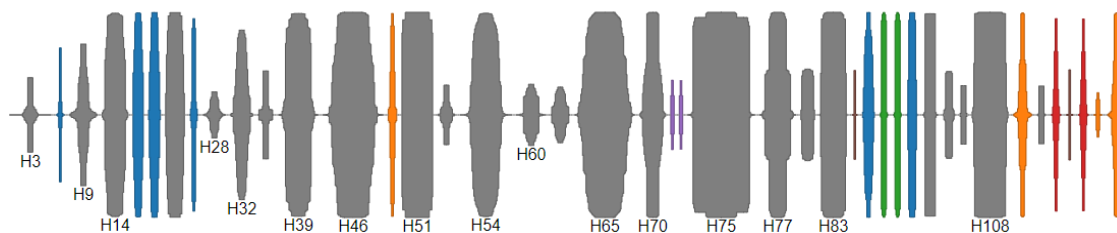
The SSEs can be depicted with or without **β-connectivity** using [Beta-connectivity ▾] . Lower arcs represent parallel connection of strands, upper arcs represent antiparallel connection.
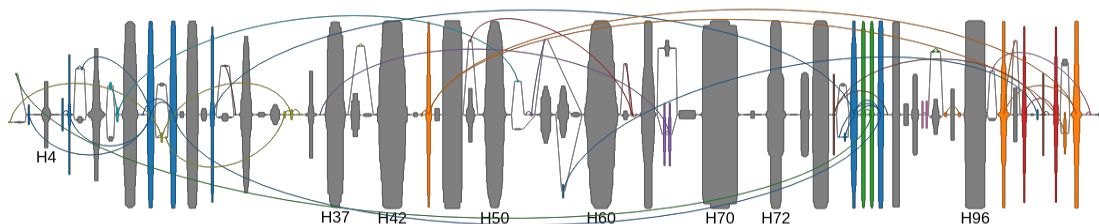
- **With β-connectivity: Default.**
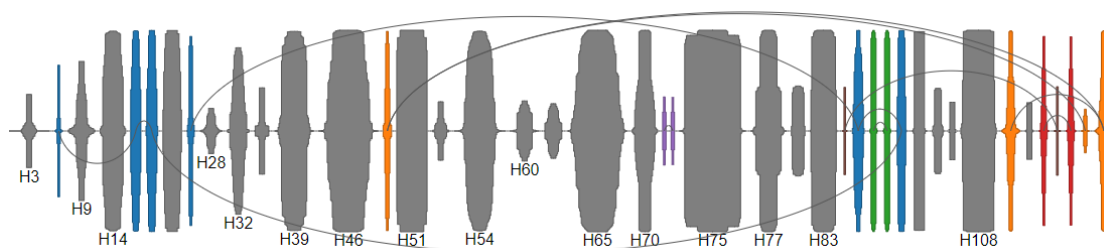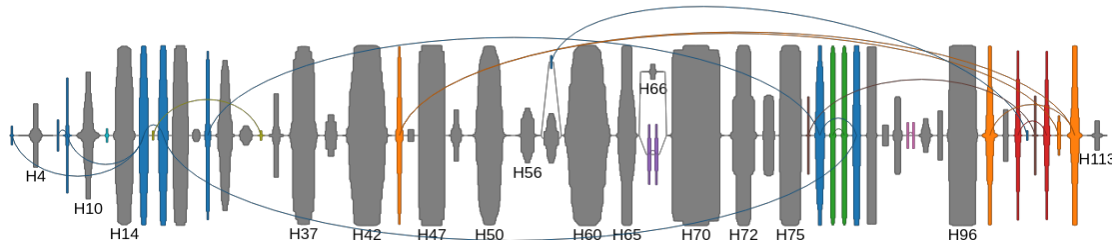

- **Without β-connectivity:**

Only SSEs with the occurrence higher than the threshold defined by [Occurrence threshold: 20% ▾] are shown, see examples:
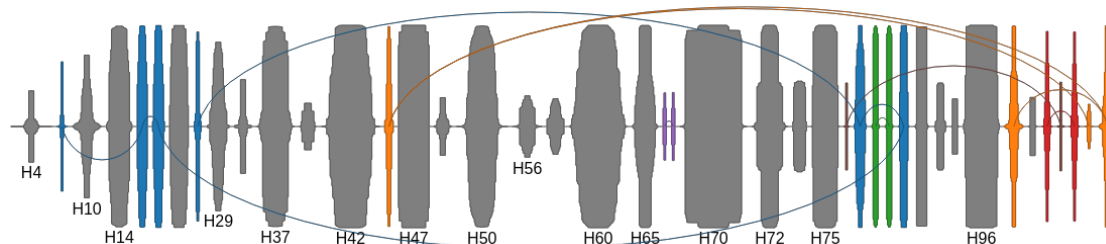
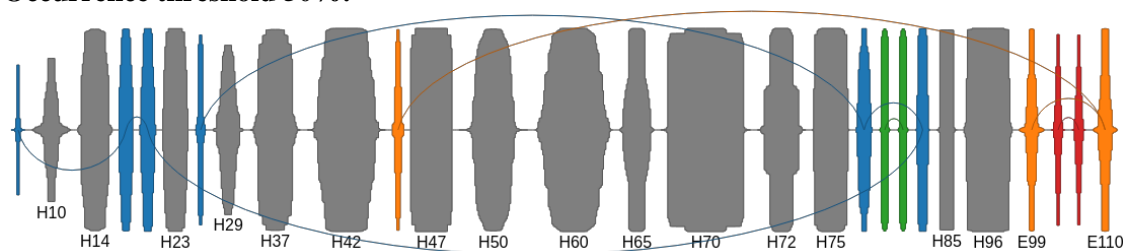- **Occurrence threshold 0%:**

- **Occurrence threshold 5%:**

- **Occurrence threshold 20%:**

- **Occurrence threshold 50%:**

With very low occurrence thresholds, the relative order of some SSEs might be undefined, therefore we can see branching in the diagram.

The current visualization can be exported to a PNG image using .

The Results page also contains a figure of 3D structure SSE consensus generated together by MAPSCI and OverProt and a figure of 2D structure SSE consensus provided by 2DProts:
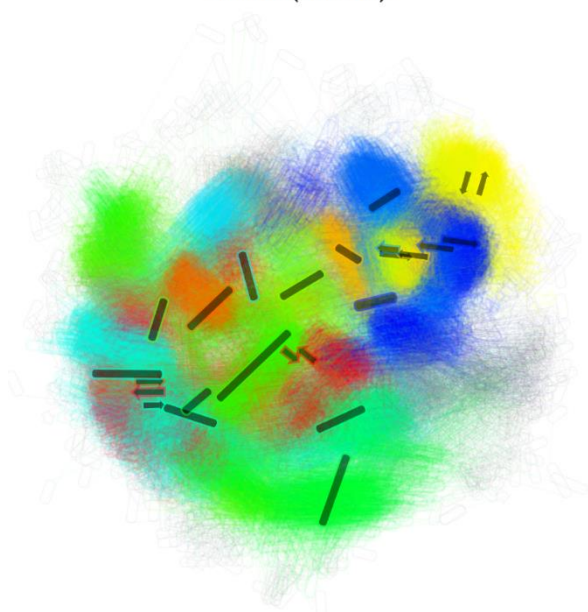
2D view (2DProts)   3D view (MAPSCI + OverProt)
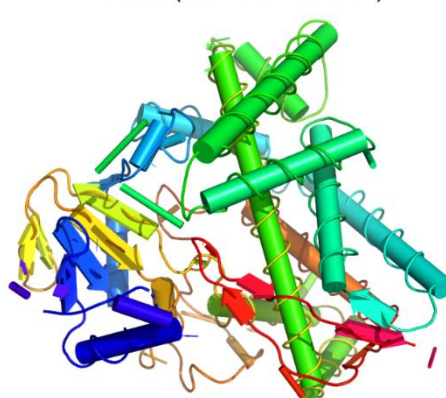
Note: If the user performed selection of input proteins via user-defined queries, the 2D structure SSE consensus is not provided, because its generation is time-demanding.

Afterwards, the Results page provides links to CATH and 2DProts pages about the protein family. Finally, Results page provides a zip file summarizing the results.

Domain view page

The user can navigate to the integrated visualization of a protein domain by:

- Clicking on the example domain on the Results page (of a precomputed CATH family):

  Example domain: 1jfbA00 ⓘ

- Selecting a domain from the list of domains for a family:

  Domains: 1638 (List) ⓘ

  Included domains: 867 (List) ⓘ

- Entering a domain ID into the search box:

  1jfbA00   Go

On the domain view page, the user can observe how the 1D information on consensus sequence of a selected family from Overprot relates to the 2D and 3D information on the selected protein domain from that family:
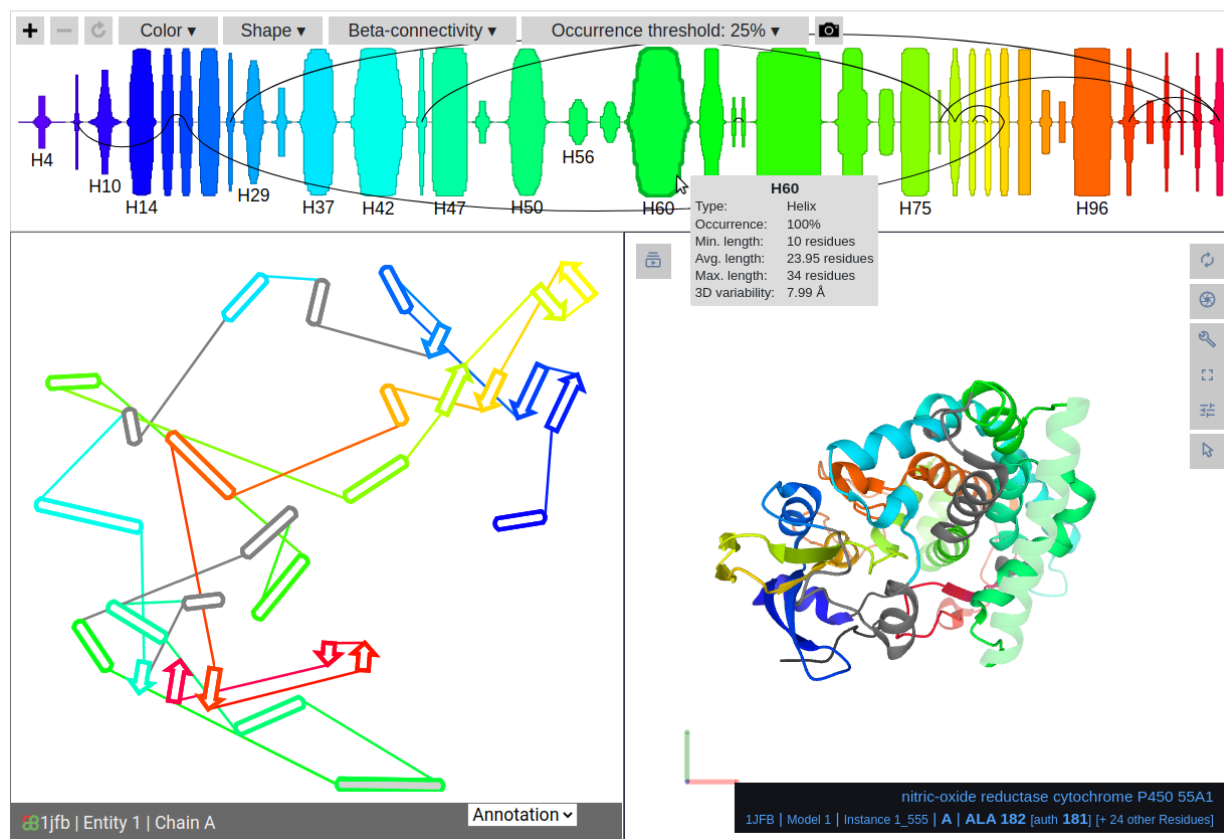
The 2D view is represented by the interactive version of 2DProts domain diagram, while the 3D view is implemented using PDBe Mol* molecular viewer.

All three tools are interconnected, so that the user can hover over an entity (an element in OverProt view, or a residue in 2DProts and PDBe Mol* view) in one view and see the corresponding entities highlighted in the other two views.

Limitations

- When a user provides CATH ID of a protein family, the consensus sequences precomputed by OverProt are used, so the results are available immediately. However, these results are based on the latest CATH release; therefore the newest protein structures are not included. From PDB entries which contain multiple domains from the same family, only one chain is included in the consensus computation.

- When a user selects the protein family members via user-defined queries, there are the following limitations:
    - No more than 500 domains can be provided.
    - The structures must be in the PDB database.
    - The user has to wait for the results (typically a few minutes).
    - OverProt is not able to recognize if the submitted domains are members of one protein family or at least in some way intercomparable proteins. Therefore, if too heterogeneous proteins are provided, the calculation might be too time-consuming and the result might be meaningless. The responsibility to provide protein domains which share structural similarity is left to the user.

- The integrated domain view may not be fully functional in some older web browsers (updating to the newest version should help). We tested it on the following browser versions:
    - Windows: Edge 85, Firefox 96, Chrome 97, Opera 83 – fully functional; Internet Explorer – no support.
    - Linux: Firefox 96, Chromium 97 – fully functional; Opera 83 – slow loading, delays in interactivity.
    - MacOS: Safari 15 – fully functional.

- 2DProts diagrams and integrated domain view are available only for precomputed CATH families (not for user-defined queries).

Note on the desktop version

Apart from the OverProt Server, the algorithm is also available in a desktop version – OverProt Core. The desktop version has some advantages, e.g., there is no hard limit on the number of processed domains, the user can provide their own structures which are not in PDB, modify the settings, and use the tool as a part of larger pipelines. The software is open-source and available at GitLab (https://gitlab.com/midlik/overprot).