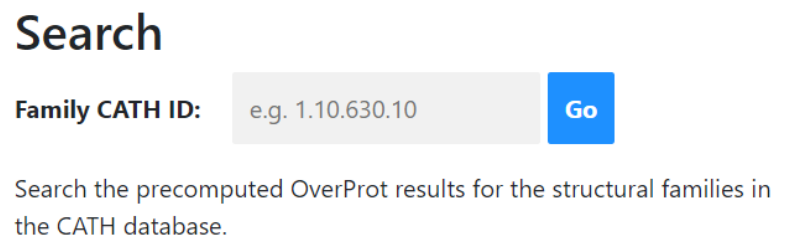
OverProt manual

OverProt (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 three pages – main introductory one (Main page), page for specification of user-defined queries (Queries page), and visualization of the results (Results page). This manual also provides information about OverProt limitations.

Contents

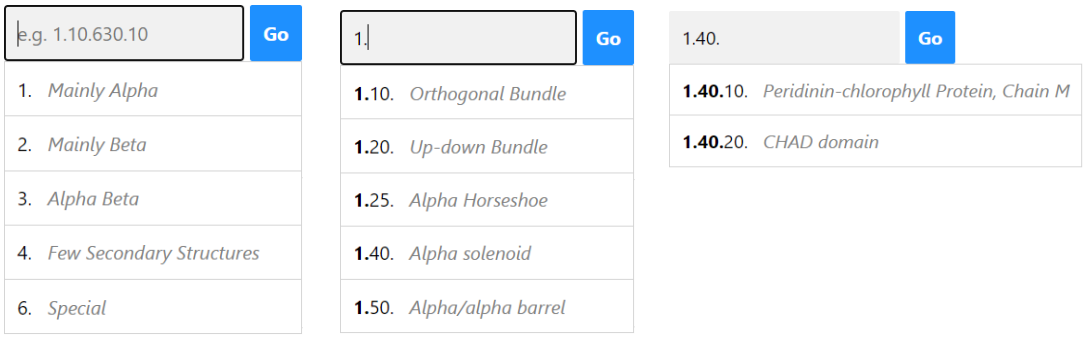
Main page

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.

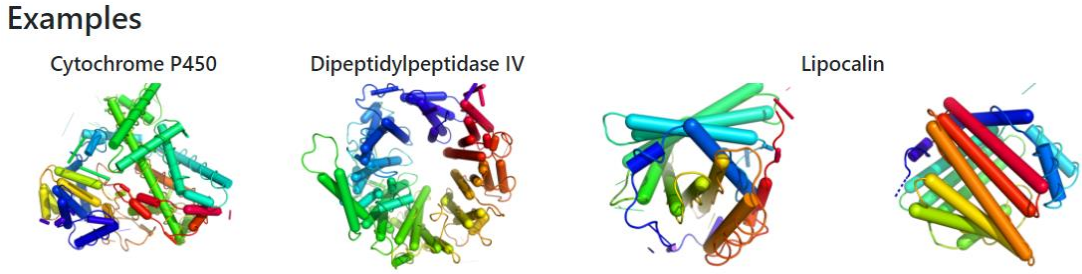
The search box also provides suggestion functionality, showing the IDs and names of existing CATH families, organized by CATH Class, Architecture and Topology (the number of suggestions is limited to 10), e.g.:



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**:

By clicking on ⇒, the user is redirected on Queries page.

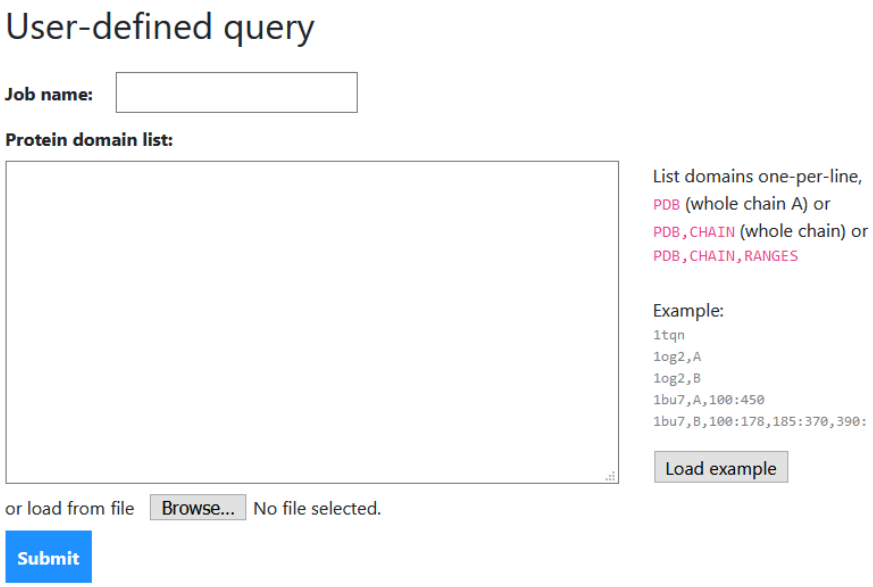
Last but not least, the Main page includes **examples**:



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:



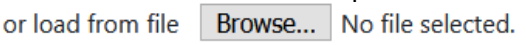
The user can make a selection of input proteins in two ways – using an **input form** or by **submitting a text file**. These ways cannot be combined.

Selection of input proteins using an **input form**:
The user can provide the following information about an input protein:

- PDB ID of a protein (will include only the chain A), for example:
  ```
  1tqn
  1OG2
  ```
- 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:
  ```

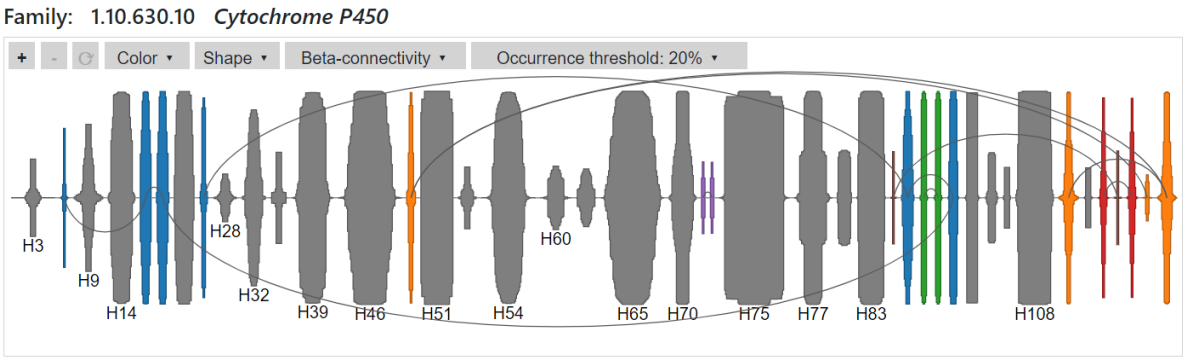
The chain IDs and residue numbers must be provided following the label_* numbering scheme (mmcif-style, corresponding to columns label_asym_id and label_seq_id in mmCIF files), not the auth_* numbering scheme (PDB-style, corresponding to columns auth_asym_id and auth_seq_id in mmCIF files and columns 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

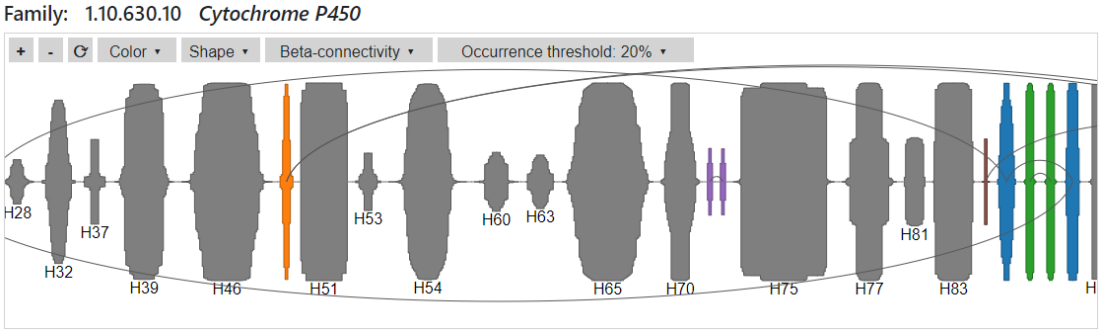
The Results page contains an interactive visualization of a SSE consensus sequence, for example:



A consensus sequence is depicted as the diagram, where:

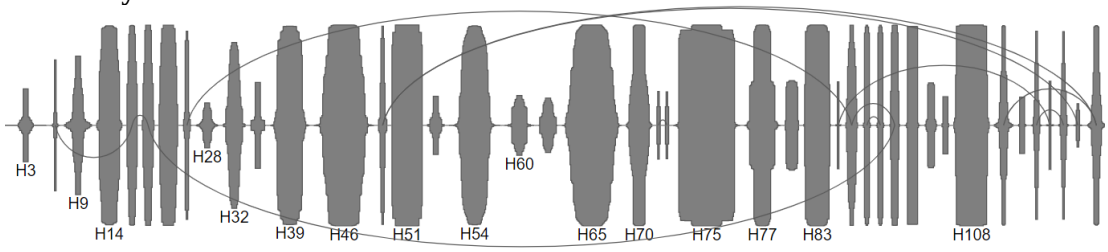
- The X-axis shows an order of SSEs, which are depicted as pure rectangles or ovals. The Y-axis indicates the occurrence. As a result, height shows the percentage of protein structures contain this SSE, while the width indicates the average length measured as the number of amino acid residues. The ovals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

Parts of the sequence can **be zoom in and out** by + -, see example here:

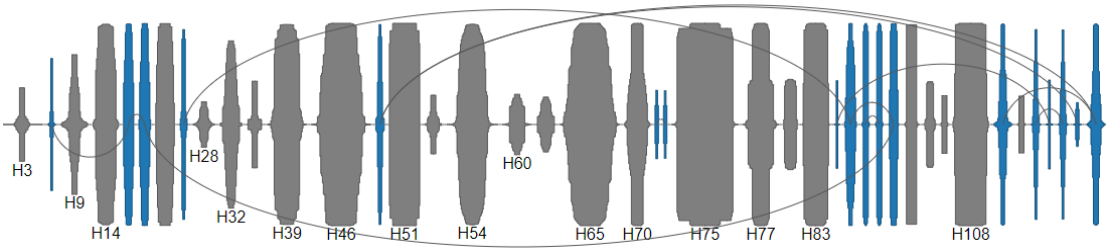


The consensus sequence can be **colored** several ways using Color, the possibilities are:

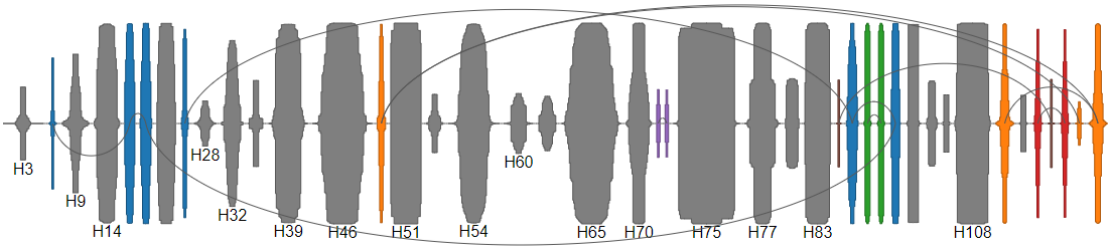
- **Uniformly:** All SSE in the same color.



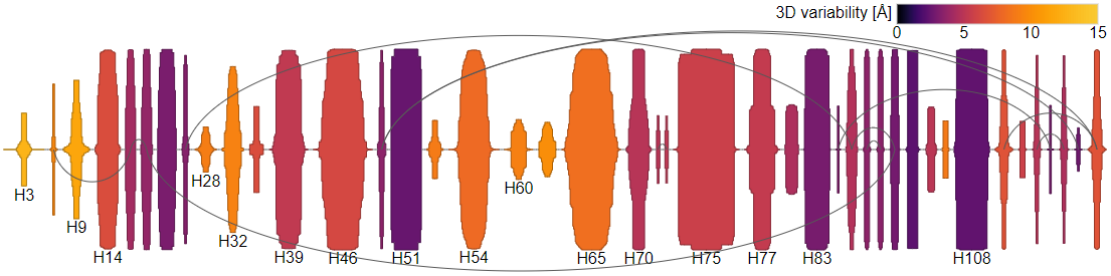
- **Type:** β-sheets in blue, helices in grey.

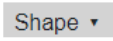


- **Sheet: Default coloring.** β-strands from the same β-sheet have the same color, while all helices are grey.



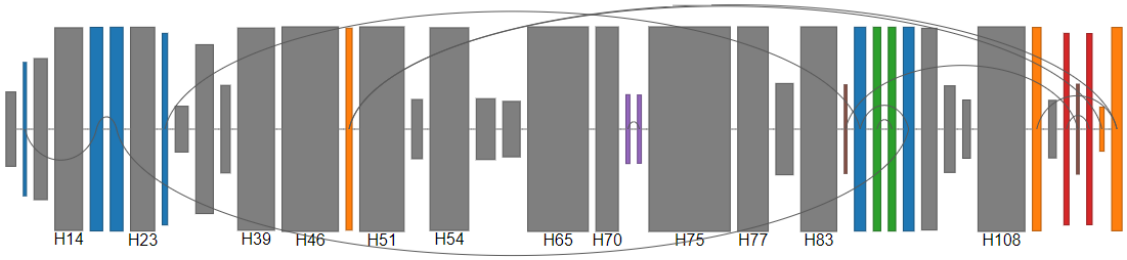
- **Variability:** 3D variability measures the standard deviation of the SSE end point coordinates. Low values (dark) indicate conserved SSE position, high values (bright) indicate variable SSE position.



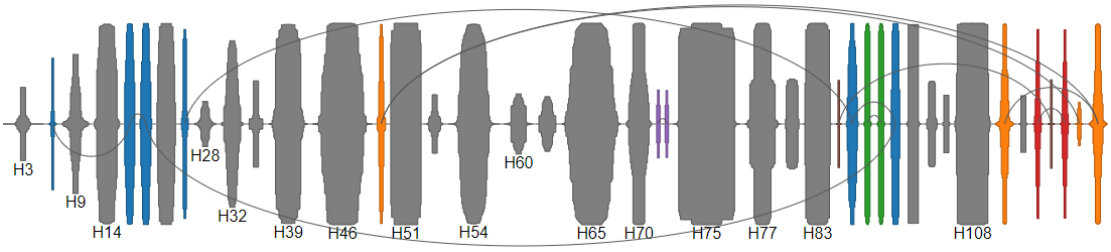
The SSEs can have two **shapes** using 

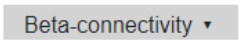
, the possibilities are:

- **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).



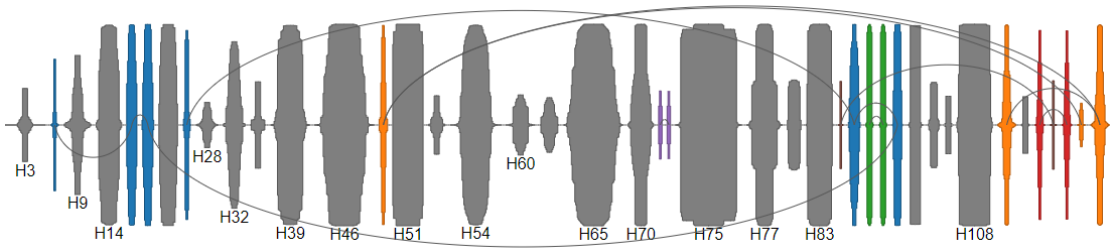
- **SymCDF: Default shape.** Show SSEs as rectangles. Height of the rectangle indicates occurrence (what percentage of structures contain this SSE), width indicates average length (number of residues).



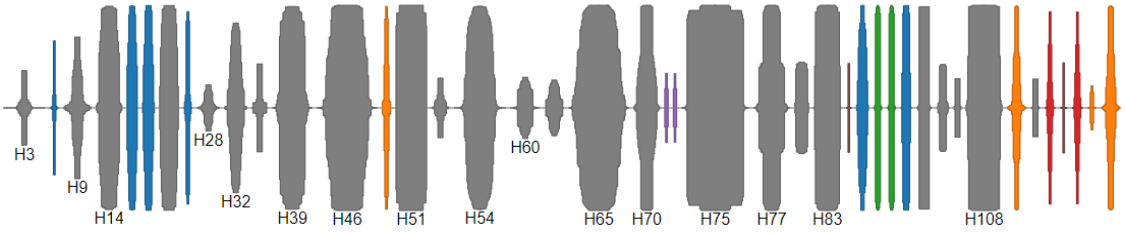
The SSEs can be depicted with or without **β-connectivity** using 

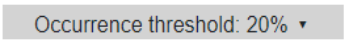
:

- **With β-connectivity: Default.**



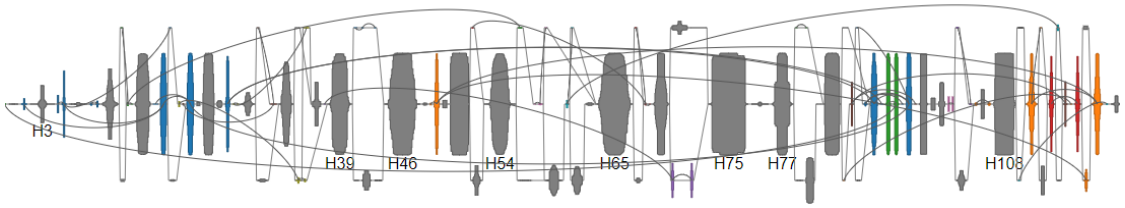
- **Without β-connectivity:**



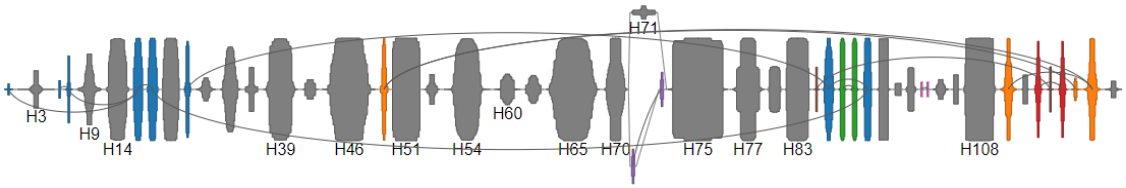
Only SSEs with an occurrence higher than a threshold defined by 

are shown, see examples:

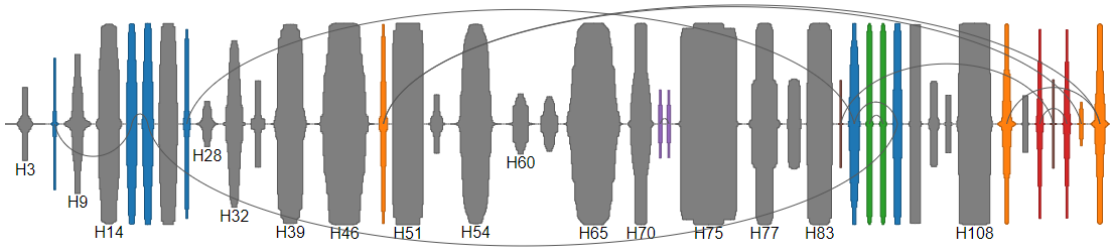
- **Occurrence threshold 0%:**



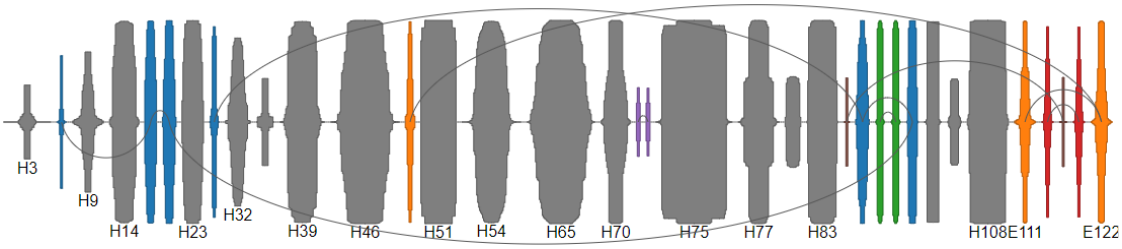
- **Occurrence threshold 10%:**



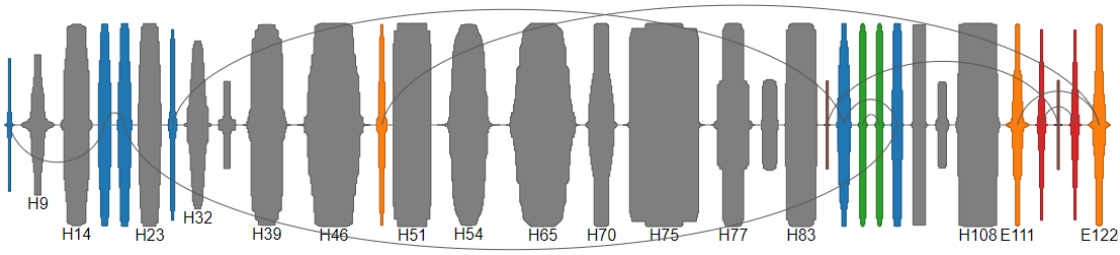
- **Occurrence threshold 20%:**



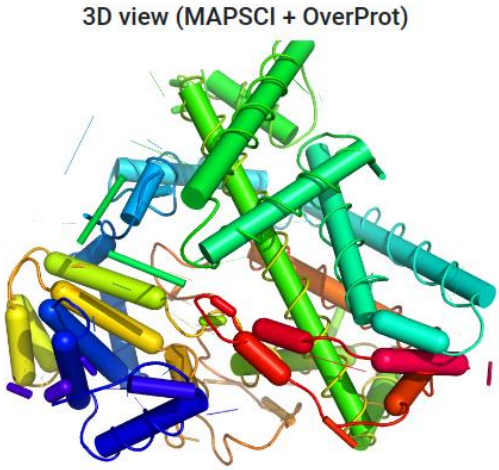
- **Occurrence threshold 30%:**



- **Occurrence threshold 40%:**



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:



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.

Limitations

- When a user provided CATH ID of a protein family, the consensus sequences precomputed by OverProt are used. The update is done once per month. It markedly saves computational time and makes the results immediately available. But in parallel, it brings the following limitations:
    - The newest protein family members are not included.
    - Only A chains used for the precomputation.
- When a user selected the protein family members via user-defined queries, there are the following limitations:
    - No more than 500 PDB IDs can be provided.
    - A user have to wait for the results.
    - OverProt is not able to recognise, if the submitted PDB id define members of one protein family or at least some way intercomparable proteins. Therefore, if too heterogeneous proteins are provided, a calculation is too time-consuming.