December 2020

About

ProteoVision is a webserver designed to visualize phylogenetic and structural information about ribosomal proteins in multiple dimensions. ProteoVision complements the previously established ribosomal RNA visualizer, RiboVision. ProteoVision was developed in the Center for Origins of Life (Georgia Tech, Atlanta GA) by Petar Penev, Loren Dean Williams and Anton S. Petrov. Please address your questions regarding ProteoVision to RiboZones@gmail.com.

Contributors

Caeden Meade, Holly M. McCann, Aparna Maddala, Chad R. Bernier, Vasanta L. Chivukula, Maria Ahmad, Aakash Sharma, Claudia Alvarez-Carreño.

Licensing

Basic Navigation

Overview

ProteoVision is operated via the main Navigation panel for selection/filtering and contains three main applets: i) Alignment Viewer for representation of multiple sequence alignments; ii) PDB topology Viewer for depiction of protein secondary structures; iii) MolStar viewer for visualization of three-dimensional structures. Additionally, non-mappable statistical data (*e.g.* amino acids frequencies) are visualized in a separate window using Plotly applet. Detailed description of all functions for ProteoVision is listed below. ProteoVision also contains an optional interactive guided tour with a brief description of each functional element.

Selecting phylogenetic group(s)

To retrieve a specified alignment from the DESIRE database, follow the steps below:

1. Click on the dropdown menu **Select a phylogenetic group**; the three dropdown nodes are labeled with the three domains of life: Eukarya, Bacteria, and Archaea.
2. Click on the left-side drop-down arrow of any phylogenetic group in the dropdown menu; the phylogenetic subgroups of that group become visible. For example, upon clicking the left-side arrow of “Bacteria,” the following groups are made visible to the user: “Bacteroidetes,” “Chlorobi,” “Cyanobacteria,” “Proteobacteria,” etc. This step is recursive.
3. Add a phylogenetic group to the list of phylogenetic groups for which the alignment will be retrieved. This list is displayed at the top of the drop-down structure. Note that when zero groups are selected this field displays the text **Select a phylogenetic group**. Multiple groups can be selected.
4. The user can also directly search for a phylogenetic group by typing in its name.

Selecting a protein alignment

Upon selection of a set of phylogenetic groups (See Selecting phylogenetic group(s)), a list of compatible protein alignments is pulled from the DESIRE database in a new dropdown menu. The list of alignments is updated after every new addition in the phylogenetic browser, the alignments list is filtered to show alignments that contain all selected phylogenetic groups. For example, if a user selects Archaea,

Bacteria and Eukarya, only universal rProtein alignments will be shown since only universal proteins are present in all three TOL branches. The user selects a protein alignment from the available ones and the alignment is displayed in the alignment viewer; the rows of the alignment are phylogenetic groups, and the columns of the alignment are arranged according to individual amino acid indices.

Some rProteins can have more than one alignment. For example, aL18 is a protein present in Archaea and Eukarya, and there are three possible alignments for it: one that includes only archaeal sequences (aL18), one that includes only Eukaryotic ones (eL18) and one that includes both (aeL18). The alignment will still be filtered to include only species selected in the phylogenetic browser. For example, if the user selects Eukarya from the phylogenetic browser and aeL18 from the alignment menu, they will fetch an alignment that was built from archaeal and eukaryotic sequences but is truncated down to show only eukaryotic sequences.

Selecting a structure for visualization and mapping

Once an alignment of a specified protein is selected, the option to select its structure for visualization and mapping becomes available. Selecting a structure is a two-step process:

1. First the user selects a 3D structure available from PDBe. Structures from the PDBe are filtered by the polymer of the selected alignment. Filtered PDB IDs are available in a dropdown menu when using the DESIRE database. When using a custom alignment, no filtering is done on PDBs and the user can write in any 4 letter PDB ID. For user convenience, the first three PDB ID options are never filtered and are those for the cytosolic ribosomes of *E. coli* (4V9D), *P. furiosus* (4V6U), and *H. sapiens* (4UG0). It is possible to encounter Eukaryotic ribosomal structures even if the user has selected a bacterial protein; these structures will be non-cytoplasmic (e.g., plastid or mitochondrial).
2. Once a PDB ID is selected, the available chains of that structure are filtered by the polymer identity present in the selected alignment. No filtering is done when using a custom alignment and all chains of the structure are available. The user can select one chain which will load the topology and MolStar viewers.

Creating a structure-alignment mapping

To ensure the match between the selected MSA and the sequence from a selected structure and to properly visualize the data, we compute an alignment between the sequences of the structure and the alignment. This is done internally on the server using the mafft program with the –addfull option (Katoh; 10.1093/molbev/mst010) and it does not require any action from the user. However, if the structural sequence has extra positions compared to the alignment, ProteoVision displays an error message located between the alignment and 2D/3D viewers. The message warns the user how many positions failed to map properly.

Selecting attribute data to map

The user has an option to map either calculated mapping data (described below) or custom data (supplied by a user) onto the 2D and 3D viewers. Available data attributes may be selected from a dropdown menu in the lower right corner of the topology viewer. There is also an option in the sidebar to upload custom mapping data. This data should be uploaded as a .csv; an example file is provided on the ProteoVision site. Once the custom data has been uploaded, it will be added as a mapping option in the dropdown menu of the PDB topology viewer applet.

Advanced features

Frequencies

Once an alignment has been selected, the user can click the **show amino-acid frequencies** button to display amino acid frequencies for that alignment. Amino acid frequencies are calculated by processing the alignment fasta using Biopython, and the resulting data is displayed as a faceted boxplot through the Plotly graphing library. Every point on an amino acid frequency plot represents the frequency of a single amino acid for a single species within the alignment. The species and amino acid frequency associated with each point can be viewed by hovering over the point.

If a user chooses from the **Select secondary structure** dropdown, the amino acid frequencies will be calculated for the specified secondary structure within the polypeptide. Users can choose to display frequencies from helix residues, coil residues, or strand residues.

Masking

Once a polymer has been selected, the user may select to **mask/unmask 2D and 3D residues**. This allows for coloration of only selected residue ranges specified by the user. All other residues will be colored in white and will have hovering functions disabled. The overall structure of the protein will still be visible.

Range selection

The user also has the option to **cut/uncut 2D and 3D residues**. This allows the user to view only the portion of the protein specified by the entered residues. The rest of the protein structure will be removed rather than colored white as in the masking feature.

Synchronization of navigation between the panels

Navigation between all panels is synchronized. Hovering over an alignment position highlights the corresponding residue in the Topology and MolStar viewers. Reversely, hovering over a residue in the Topology or MolStar viewers highlights the alignment and the other structural viewer. When the amino-acid frequencies are shown, hovering over the datapoints on the graph highlights the current species in the alignment viewer.

Guide

ProteoVision offers an interactive guide that demonstrates the steps a user can take to fully utilize ProteoVision capabilities. The guide always starts on the users first visit and can be launched at any time with the **Help** button. The user can step through the guide using the keyboard arrow keys or the **Next** and **Previous** buttons on the guide pop-ups. Ending the guide with **Skip tour** button will erase the current session and reset the viewports.

User upload mode

ProteoVision supports the upload of custom alignments. The user can upload any fasta format alignment through the **User upload** menu and calculate amino-acid frequencies and mapping data from it. The user can select a PDB ID to visualize a structure and map the calculated data from their alignment.

Selecting a structure for visualization and mapping in user upload mode

After uploading an alignment, ProteoVision will use the first sequence of the alignment to perform a BLAST search of the PDB database (https://www.ebi.ac.uk/Tools/common/tools/help). BLAST results are filtered by E-value lower than 10^{-5} and are used to populate a dropdown menu in the PDB input field. The

dropdown menu is searchable, and shows filtered results depending on the user input. After a BLAST is complete the polymers associated with the polymer selection box will be filtered by the BLAST results.

A BLAST search can take a long time, so ProteoVision displays a message that BLAST is running under the PDB input field ("BLASTing available PDBs"). While waiting the user can input any 4 letter PDB ID in the PDB input field the fetched polymers for that PDB ID will not be filtered. When the BLAST search is complete the message will change to indicate completion ("Completed BLAST for similar PDBs").

DESIRE-API

ProteoVision uses an API service which provides data about rProtein nomenclature, sequences, alignments, and annotations. Furthermore, it provides data about species phylogeny. The API is available at https://proteovision.chemistry.gatech.edu/desire-api/.

Saving

Saving the alignment (fasta & image)

The alignment retrieved from the DESIRE database can be downloaded in fasta format with the **Download alignment** button. The current viewport of the alignment can also be saved as a png image with the **Download alignment image** button. Only the currently viewable region of the alignment will be saved in this way.

Saving secondary structure image (svg)

The secondary structure image may be downloaded as a .svg file by clicking the "**S**" button in the bottom right corner of the topology viewer.

Saving 3D structure image (png)

The image of the 3D structure may be downloaded as a .png file or viewed in browser by clicking the **Screenshot/State** button, which appears as a wheel in the upper right corner of the 3D viewer. Upon clicking this button, the user may choose to either keep the default white background or make the background transparent by turning transparency to **On**. They may also choose to either include or exclude 3D axes which show the orientation of the protein.

Saving computed data (csv)

Mapping data (described below) calculated for a given alignment by the webserver can be saved in .csv format from the **Download mapped data** button. The first column of the saved file is labeled Index and indicates the residue number. The rest of the columns hold the calculated attributes for each alignment column mapped on the current structure indices.

Saving frequencies (png)

The amino acid frequency plot can be saved as a .png file by hovering over the image and clicking on the **camera icon** in the top center.

Saving a ProteoVision session

At any point, the user can save their progress with the **Save session** button. This will download a .json file that holds information about the currently loaded alignment, structure, and custom mapping data. The session file does not save information about the masking and truncation ranges.

ProteoVision Data

Phylogeny (SEREB)

The subset of 152 species from the SEREB (Sparse and Efficient Representation of Extant Biology, https://doi.org/10.1093/molbev/msy101) database was organized into a phylogenetic browser using a tree topology from the Banfiled lab (https://doi.org/10.1038/nmicrobiol.2016.48)

Alignments

Each ribosomal protein has an associated MSA. First, an MSA reference was generated with MATRAS (https://doi.org/10.1093/nar/gkg581) from multiple structure superimpositions. Then, amino acid sequences of species from the SEREB database were added to the reference alignment using MAFFT (https://doi.org/10.1093/bioinformatics/bts578).

2D maps

Topologies of the protein secondary structures (Laskowski 10.1093/nar/gkn860) were exported into PDB topology viewer as APIs provided by EMBL-EBI PDBe and available at https://www.ebi.ac.uk/pdbe/api/doc/.

3D Structures

3D structures were fetched from the PDBe using the APIs of EMBL-EBI coordinate server https://www.ebi.ac.uk/pdbe/coordinates/. The selection of ranges was implemented using the syntax of the LiteMol's coordinate server https://coords.litemol.org/

Alignment associated data (Fold, Phase)

DESIRE holds annotations of domain architecture from ECOD (Cheng; 10.1371/journal.pcbi.1003926) and ribosomal phase definitions (Kovacs; 10.1093/molbev/msx086) at residue level for one representative species (*E. coli*). Using the alignments, ProteoVision retrieves these annotations for each column of an alignment and displays them as a hovering pop-up next to each residue.

Available attributes for calculated mapping data:

Amino Acid frequencies

Amino acid frequencies in each column of an MSA were adjusted for presence of gaps. Thus, the gap frequencies were prorated and were treated as a uniform distribution among all possible amino acid characters, such that a single character in a gap counts as 0.05, as described by Bernier et al. (10.1093/molbev/msy101).

Shannon Entropy

The Shannon entropy (as well as all properties listed below) was computed from the gap adjusted probabilities as:

$$H_{SE}(n) = -\sum_{i=1}^{c} p_i(n)\log_2 p_i(n) \;\cong -\sum_{i=1}^{c} f_i(n)\log_2 f_i(n)$$

Two group comparison (TwinCons)

In case of two groups selected in the phylogeny browser, ProteoVision provides an additional option to compute an in house developed score, TwinCons. TwinCons is computed for a single position of the MSA that compares two pre-defined groups (represented by vectors of the gap adjusted amino acid

frequencies) based on their similarity defined by the pre-computed substitution matrix. TwinCons represents the transformation price between the two vector columns related by the substitution matrix.

Charge, hydropathy, hydrophobicity, polarity, mutability

The physico-chemical properties for each position within an MSA are computed as average properties for a given distribution of the amino acid frequencies. The tabulated values for each property were obtained from the available literature: i) charges (10.1186/1758-2946-5-39); ii) hydropathy (10.1016/0022-2836(82)90515-0); iii) hydrophobicity (10.1093/protein/5.5.373); iv) polarity (10.1016/0022-5193(68)90069-6); v) mutability (10.1093/bioinformatics/8.3.275).

Color Schemes

Each calculated attribute is mapped on a matplotlib colorscheme. For single continuum attributes (like Shannon entropy or Polarity), ProteoVision uses single continuum colormaps like plasma and viridis. For diverging data attributes (like Charge or TwinCons), ProteoVision uses diverging colormaps like Blue-White-Red or Green-White-Purple. All colormaps were generated with the python matplotlib library and exported to JavaScript with the js-colormaps package (https://github.com/timothygebhard/js-colormaps). Further information about colormaps in matplotlib (https://bids.github.io/colormap/)

Import User supplied Data

Import an external dataset for a pre-selected alignment

Once an alignment and a structure have been selected, the option becomes available to upload custom data for mapping onto the selected structure for visualization in the topology and MolStar viewers. The data should be supplied in a .csv (comma-separated values) file format. The first row of the csv file contains the headers for each column. An Index column should always be indicated. All other additional columns require a unique header definition. The Index column has the residue number to which the user-supplied data will be mapped. The rest of the columns have the data that will be mapped in the form of numerical values. Once a correct csv file has been uploaded, the selected structure will be colored in the topology and MolStar viewers according to the values in the csv file. Different columns in the csv file will appear as different **Annotations** in the dropdown menu in the lower right corner of the topology viewer.

Upload an external multiple sequence alignment.

The **User Upload** mode allows the user to use all the ProteoVision features with an external MSA. An alignment in a .fasta file format must be selected and then uploaded with the **Upload alignment** button. Once an alignment is uploaded, it will be displayed in the Alignment viewer. The steps for structure selection, mapping, attribute calculation, and saving are the same as previously described.

Import a saved ProteoVision session file

At any time, the user can upload a previously saved ProteoVision session file with the **Load session** button. This will restore the progress at time of saving for selected/uploaded alignment, calculated/uploaded data attributes, and selected structures in the 2D and 3D viewers. Currently the sessions do not recover masking or truncation ranges.

Acknowledgments:

This work was funded by the National Aeronautics and Space Administration grant 80NSSC18K1139 awarded to LDW and ASP.

We would like to thank the following collaborators, whose input has been invaluable to our work:

- Prof. Khanh Dao Duc & Mr. Artem Kushner at University of British Columbia (https://kdaoduc.com/)
- Mr. Rohan Gupta at Georgia Institute of Technology
- The group of George Fox at University of Houston
- Dr. Burak Gulen, Postdoctoral Scientist, University Medical Center Hamburg Eppendorf
- Ms. Yulia Dumov, Sr. UX/UI Designer, Tufts Technology Services (TTS)

Our webserver would not be possible without these free resources:

- Sequences and taxonomies were retrieved from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) and UniProt (https://www.uniprot.org/).
- rProtein structural annotation was retrieved from ECOD (10.1371/journal.pcbi.1003926; 10.1002/prot.24818) (http://prodata.swmed.edu/ecod/).
- Alignment viewer – modified version of the react implementation for the MSAViewer (10.1093/bioinformatics/btw474) (https://github.com/plotly/react-msa-viewer).
- Topology Viewer – modified version of the PDBe topology viewer (https://github.com/PDBeurope/pdb-topology-viewer).
- PDB topologies provided through the EBI PDBe REST API (https://www.ebi.ac.uk/pdbe/api/doc/topology.html).
- MolStar Viewer – (https://molstar.org/viewer/; https://github.com/molstar/molstar).
- Parsing of sequences and alignments and calculation of frequencies was possible thanks to the Biopython package (https://biopython.org/).