*Nucleic Acids Research, 2021* **1**
*doi: 10.1093/nar/gkab351*

# ProteoVision: web server for advanced visualization of ribosomal proteins

**Petar I. Penev [1,2], Holly M. McCann[2], Caeden D. Meade[2], Claudia Alvarez-Carreño [1,3], Aparna Maddala[2], Chad R. Bernier[2,3], Vasanta L. Chivukula[1,2], Maria Ahmad[2], Burak Gulen[3], Aakash Sharma[2], Loren Dean Williams [1,2,3] and Anton S. Petrov [1,2,3,*]**

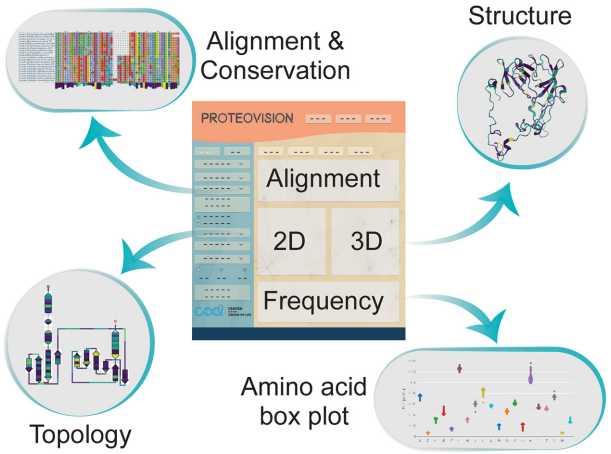
[1]NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA, [2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA and [3]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received March 09, 2021; Revised April 11, 2021; Editorial Decision April 21, 2021; Accepted April 21, 2021

## ABSTRACT

**ProteoVision is a web server designed to explore protein structure and evolution through simultaneous visualization of multiple sequence alignments, topology diagrams and 3D structures. Starting with a multiple sequence alignment, ProteoVision computes conservation scores and a variety of physicochemical properties and simultaneously maps and visualizes alignments and other data on multiple levels of representation. The web server calculates and displays frequencies of amino acids. ProteoVision is optimized for ribosomal proteins but is applicable to analysis of any protein. ProteoVision handles internally generated and user uploaded alignments and connects them with a selected structure, found in the PDB or uploaded by the user. It can generate de novo topology diagrams from three-dimensional structures. All displayed data is interactive and can be saved in various formats as publication quality images or external datasets or PyMol Scripts. ProteoVision enables detailed study of protein fragments defined by Evolutionary Classification of protein Domains (ECOD) classification. ProteoVision is available at http://proteovision.chemistry.gatech.edu/.**

## GRAPHICAL ABSTRACT




## INTRODUCTION

Proteins commonly attain function upon assuming ordered (folded) structures. In the early 1950s, when Pauling *et al.* (1,2) proposed secondary structure elements (α-helices and β-sheets) and Kendrew (3) and Perutz (4) decoded the first structures of proteins, it became evident that the complexity of even a small protein is immense. It is simply impossible for a human to understand a protein from inspection of atomic positions or primary sequence. Important analytical advances have reduced complexity while capturing essential information. Some of these advances were the multiple sequence alignment by Sussman (5), and the topology representation (6) and the 3D ribbon representation (7), both by Jane Richardson. These reduced representations allow efficient visualization and evaluation of primary, secondary, tertiary and quaternary information and analysis of variety of phenomena ranging from ligand binding to complex protein folds. In this report, we present a web server that allows user-controlled mapping of information

*To whom correspondence should be addressed. Tel: +1 404 894 8338; Fax: +1 404 894 7452; Email: anton.petrov@biology.gatech.edu

from multiple sequence alignments and variety of other sequence and structural sources simultaneously on topology and ribbon representations. We have integrated analysis of sequence space and structure space, abolishing barriers to instant mapping of data between them.

Our initial application of this technology is to ribosomal proteins. The ribosome is a universal component of life found in every living cell that produces every protein in extant biology. The ribosome itself is a complex, data-rich macromolecular assembly that consists of several large ribosomal RNA molecules (rRNAs) and over fifty ribosomal proteins (rProteins). Structural analysis of ribosomes has been pioneered by the works of A. Yonath (8,9), T. Steitz (10), V. Ramakrishnan (11), and J. Frank (12). Ribosomes are exceedingly complex in terms of structure and function, widely distributed in biological systems, and extremely dense in structural and historical information (13,14). Recent efforts in ribosomal structure determination (15,16) and sequence analysis (17–19) have uncovered additional information about ribosomal protein evolution and phylogeny (20–22). Yet, due to their complexity, centrality, and universality in biology (23), ribosomes pose a special set of problems and opportunities for visualization and analysis. Which elements of the ribosomal proteins are conserved in structure? Are they also conserved in their sequence? What is the difference between a given ribosomal protein in archaea and bacteria? It is challenging to find the answers to these questions by visualizing a single data set (an alignment or a 3D structure). It is far more challenging to synchronize and perceive the information from multiple sources.

Advances in computer technology and the explosive increase in the number of macromolecular structures have triggered the development of tools for molecular visualization (24). Some of these tools have been incorporated in web applications that portray a single molecular representation, visualize a single dataset, or perform a single type of analysis (25). Currently there is a shift towards simultaneous display of multiple data types (26) and interactive crosstalk between multiple visualizations providing great benefits to the scientific community (27). We present the ProteoVision web server for sequence and structure visualization of rProteins, available at https://proteovision.chemistry.gatech.edu. ProteoVision enables users to simultaneously visualize information related to ribosomal proteins across phylogeny at levels of primary, secondary, and three-dimensional structure. Additionally, ProteoVision supports custom alignments for any protein; its integration with available 3D structures and topology diagrams, as well as custom structure and data upload options make it a general tool for visualization and data mapping. We believe ProteoVision is the first readily accessible web server for instantaneous mapping of user data simultaneously onto alignment, topology diagrams and three-dimensional structures.

## WEB SERVER DESCRIPTION

ProteoVision is a web server designed to facilitate comparative studies of proteins in 1D, 2D and 3D. ProteoVision has two modes of operation: (i) the DESIRE mode provides rProtein alignments from the DESIRE database and (ii) the User upload mode enables the upload of external alignments and user-supplied 3D structures. ProteoVision is operated via a Main Navigation panel and displays results in four synchronized applets for the alignment, topology, 3D structure, and amino acid frequency of the protein. ProteoVision directly connects an alignment and a structure by generating an alignment-structure mapping. ProteoVision calculates conservation and physicochemical properties from a selected alignment, which can be mapped onto the structure. ProteoVision has an optional interactive guided tour with a description of each functional element.

### Web server development

ProteoVision is a web server hosted at Georgia Institute of Technology and served by Apache/RHEL7. The web server was developed using Python 3.7.3 on WSL. The front end development was performed with HTML5/JavaScript using React and Vue frameworks. The back end was designed using Django 2.2.13 and a MySQL database. The front and back ends are decoupled through a Django REST framework (Figure 1). ProteoVision also relies on information provided by several external APIs and databases.

### Infrastructure

A principal component of the ProteoVision server is the DESIRE database (DatabasE for Studying and Imaging of Ribosomal Evolution). This relational MySQL database connects information on species taxonomy, gene annotation, polymer sequences, and ribosomal protein nomenclature with polymer alignment indices. The back end parses user queries to generate a multiple sequence alignment for a specified taxonomy group and polymer type. Additional structural (28) and evolutionary (29) annotations are retrieved and served to the front end. The front end integrates the Taxonomy Browser within the Main Navigation panel with the Viewers panel containing four synchronized visualization applets: (i) MSAViewer (30) for portrayal of multiple sequence alignments; (ii) PDB Topology Viewer for depiction of protein topology diagrams (27,31); (iii) Mol* viewer (32) for visualization of three-dimensional structure representations, iv) Frequency box plot for interactive visualization of amino acid abundances within the multiple sequence alignment (Figure 2). ProteoVision provides navigation for session management, API service for the DESIRE database, detailed documentation (accessible via 'About' button) and an interactive guided tour ('Help').

### Main Navigation panel

Navigation in ProteoVision is governed by the Main Navigation panel, which consists of the Navigation Mode switch, the Taxonomy Browser, and the alignment upload or selection menus. Additionally, the Main Navigation panel provides an option to select a single protein domain based on Evolutionary Classification of protein Domains (ECOD) (28,33) or a user specified range for a given protein chain. The Main Navigation panel also enables visualization of the amino acid frequencies computed for a selected alignment.

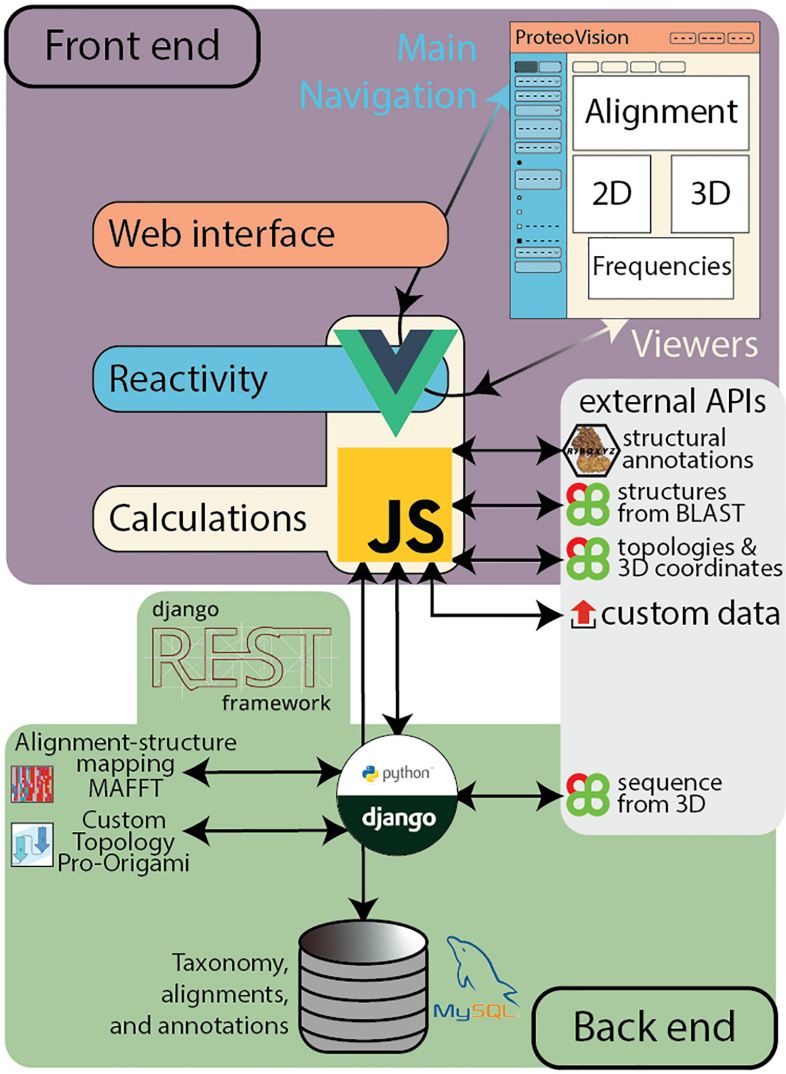



**Figure 1.** Architecture of the ProteoVision web server. ProteoVision code is decoupled between the front end and back end. In the back end (green), the Django web framework queries and serves a REST API of the MySQL database. The back end also executes alignment-structure mapping and queries the EBI server for sequences from selected 3D structures. In the front end (purple), an integrated Vue framework with JavaScript (tan) builds the Main Navigation panel (blue), provides interactivity, calculates conservation and physicochemical properties, queries the back end and external APIs, and displays the viewers (tan).

*Navigation modes.* ProteoVision operates in two modes which differ in their input datasets and protocols. Once the inputs are completed, both modes converge to the same framework and share similar functionality.

***DESIRE mode.*** ProteoVision provides a collection of rProtein alignments stored in the DESIRE database. The database includes 147 structure-guided and manually curated rProtein alignments. These alignments contain 9448 polymer sequences from 179 species that sparsely and efficiently sample the tree of life (34), include recently discovered phylum of Asgard species (35,36), and provide a comprehensive sampling for eukaryotic species (37). The list of species in the DESIRE database and their taxonomic identifiers are provided in the ProteoVision documentation and Supplementary Dataset S1. The alignments are dynamically constructed using the Taxonomy Browser and displayed in MSAViewer. The DESIRE mode provides a quick integration of an rProtein alignment with 3D structure and its topology diagram by taking advantage of annotations of ribosomal proteins from https://ribosome.xyz/ (riboXYZ) and REST APIs from PDBe and resolving the mismatches in naming of ribosomal proteins to the modern notation (38).

***User upload mode.*** ProteoVision is designed as a general purpose resource and is not limited to visualization of ribosomal proteins. ProteoVision provides an option to upload and visualize an alignment of any protein and connect it with a 3D structure and corresponding topology diagram if those are available. Once an alignment is uploaded, it will be displayed in the MSAViewer and calculated conservation will be displayed as a bar graph under the alignment. ProteoVision further enables a generic mechanism of inte-

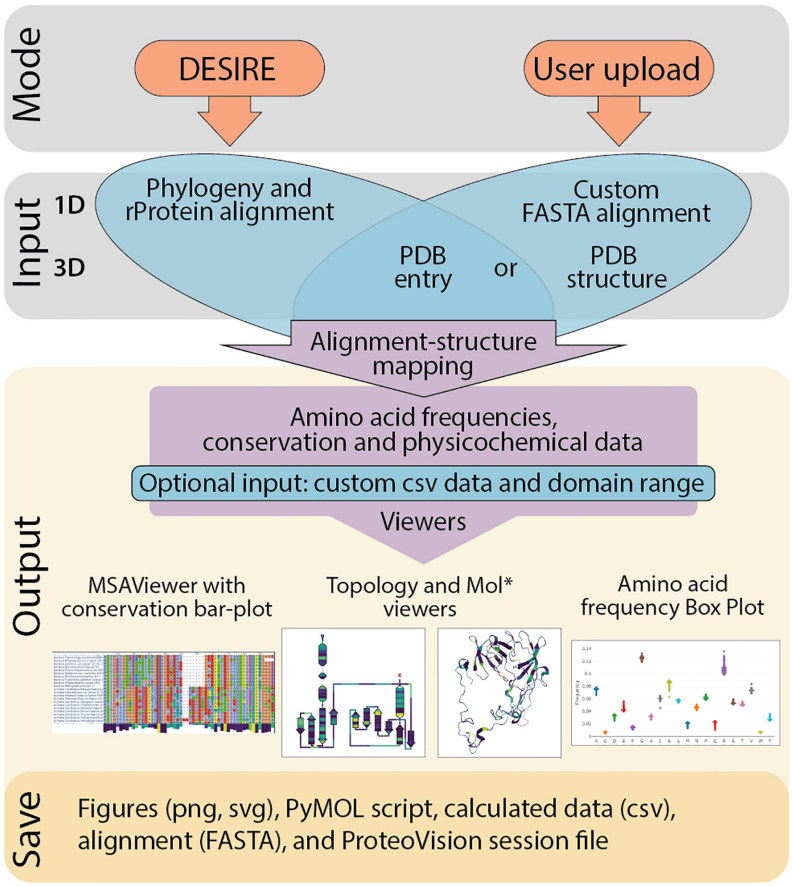



**Figure 2.** Workflow of the ProteoVision web server. ProteoVision operates in two modes, depicted in orange ovals. After selecting or uploading an alignment and a PDB entry (blue), ProteoVision creates the alignment-structure mapping and computes conservation and physicochemical properties from the alignment (purple). ProteoVision displays the four viewers (tan) and provides options to input custom CSV data or domain ranges for truncation and masking (blue). Finally, ProteoVision downloads computed data and images (yellow).

grating the uploaded alignment with 3D structures and the topology diagrams.

*Taxonomy browser.* The Taxonomy Browser provides navigation across species within the DESIRE database. The root level consists of three groups: Bacteria, Archaea, and Eukarya. Within each group, multiple subgroups representing various taxonomic levels can be iteratively expanded. The taxonomic organization of species was adapted from NCBI (39). One or multiple groups from any level can be simultaneously selected or searched for by typing in its name. Selection of any parent group will include all species within that group available in the DESIRE database. Due to variability of rProteins in different domains of life (38,40), ProteoVision dynamically loads and filters a list of available alignments based on selections within the Taxonomy Browser. For example, if Archaea, Bacteria and Eukarya are selected, only universal rProtein alignments are shown. If Bacteria is deselected, a subset of rProtein alignments common to archaeal and eukaryotic domains is added to the alignment list.

*Alignment-Structure mapping.* A key feature of ProteoVision is its ability to connect a protein alignment with a 3D structure and topology diagram. This alignment-structure connection is done through several steps:

First, ProteoVision identifies relevant PDB structures for the sequences in the selected alignment. In DESIRE mode the identification is performed via filtering a specified rProtein within available ribosomal structures using riboXYZ API; the filtered PDB IDs are provided in alphabetical order along with species names in a searchable menu. Three ribosomal structures, representing the three domains of life, are always appended at the top of the structure list for the user's convenience. In User upload mode, the relevant PDB structures are identified through a BLAST search within the RCSB database using the first sequence of the alignment as a query. The filtered PDB IDs are sorted by BLAST results and provided in a searchable menu. In both modes the user can also input any desired PDB ID, not present in the filtered results.

Second, ProteoVision searches for relevant chains within the selected PDB ID to match those in the current alignment. A given structure may contain multiple chains of proteins, most of which are not relevant to the supplied alignment. In DESIRE mode, rProtein chains matching the specified alignment are detected by annotations from PDBe. In User upload mode, only chains with an *E*-value below $10^{-5}$ and query coverage greater than 75% from the previous

BLAST search are displayed in descending order of probable similarity. *E*-value and coverage are shown as a tooltip when hovering over a structure of interest in the 'Select/type PDB entry' menu. If an external 3D structure is provided instead using the 'Upload a custom PDB' option, ProteoVision takes it as an input for the Mol* Viewer, and generates a custom topology diagram using the Pro-Origami program (41,42).

Finally, ProteoVision connects the selected alignment and structure by coupling the alignment indices with the residue indices from the structure. To achieve this match, the web server automatically adds the sequence of a currently selected structure to the alignment using the mafft program with the '–addfull' option (43). The added sequence is only used to create a proper mapping. The reference sequence extracted from the 3D structure is appended at the top of the alignment but it is excluded from calculations of various scores. As a result of these steps, a selected or uploaded alignment should be linked to a representative 3D structure and its topology diagram, so that mapping of alignment-calculated data onto structure is possible. Each protein in the DESIRE mode can be visualized within the context of the entire ribosome upon checking 'Show ribosomal context in 3D' box.

*Range selections.* ProteoVision supports visualization of specified fragments of proteins and masking of data mapped onto topology diagrams and 3D representations. A selection of a protein fragment (using ECOD domain annotations or by a user specified range) hides the structure with mapped data outside the selected range. Masking hides the selected data outside of the specified range but preserves the display of the entire structure.

*Amino acid frequencies.* ProteoVision can display amino acid frequencies and their mean values for species within the alignment as an interactive Frequency box plot. The default frequency calculation includes all residues within the alignment. The advanced frequency calculation allows the users to select a subset of residues based on (i) secondary structures elements (helix, sheet or coil), (ii) ECOD domains or (iii) a user specified range.

*Custom data.* ProteoVision provides an option to map custom data in CSV format onto selected structure for visualization in the MSA, Topology, and Mol* Viewers. Upon uploading the file, the web server converts the data associated with protein residues to color codes based on the minimum and maximum data values. The colors are mapped onto the corresponding positions of structures in Topology and Mol* Viewers as well as the bar-plot of the MSAViewer. If multiple data sets are supplied in a single file, they will appear as different entries in the dropdown menus of MSAViewer and Topology Viewer.

### Viewers panel

*MSAViewer.* To visualize the selected or supplied alignment, ProteoVision employs a REACT implementation of the MSAViewer (30) (https://github.com/plotly/react-msa-viewer/). MSAViewer is a powerful open-source JavaScript component designed for web applications, which provides a graphical display of any MSA. Simple interactive navigation is complemented by the ability to select, color, or highlight desired alignment fragments. MSAViewer code was adapted for ProteoVision to provide integration with other plugins via (i) sharing data mappings, (ii) dynamic positioning and (iii) highlighting of a specific position within the alignment by hover events. Additionally, the bar-plot of the MSAViewer was adjusted to display conservation calculated as Shannon Entropy and to color each bar according to selected alignment-derived data. Visualizations from the MSAViewer can be saved in PNG format.

*Topology viewer.* To visualize protein topology diagrams, ProteoVision employs an adapted version of the PDB Topology Viewer (27), which is an open source JavaScript plugin of PDBe Component Library (https://github.com/PDBeurope/pdb-topology-viewer). Topology Viewer displays two-dimensional topology diagrams for proteins developed by PDBeSum (31), and provides interactive selections of protein residues and coloring capabilities. The code was adjusted to enable mapping of a variety of phylogenetic and physicochemical properties by using color gradients (https://github.com/timothygebhard/js-colormaps). ProteoVision also expands the functionality of the original plugin by allowing users (i) to map their own data; (ii) select a specific protein domain or arbitrary residue range within a specified protein chain; (iii) mask out data visualizations within a specific region, while keeping the entire structure visible for reference; (iv) save the online visualization into an external SVG file. All highlighting, coloring and range selections for the Topology Viewer on the ProteoVision site are synchronized with the Mol* viewer and the MSAViewer.

*Mol* viewer.* Visualization of 3D structures is performed via Mol* JavaScript component (32), an open-source project maintained by Protein Data Bank in Europe (PDBe) and RCSB Protein Data Bank (PDB) (44). The Mol* component was integrated with PDB Topology Viewer by calling available layout and display methods. Mouse custom events were registered to provide synchronous interactions among the Viewers. 3D structure retrieval and selection of ranges was implemented using syntax from the LiteMol coordinate server (https://coords.litemol.org/).

*Frequency box plot.* The visualization of amino acid frequencies was implemented using the Plotly JavaScript library. Upon selection of the 'show amino acid frequencies' option, a Plotly box plot displays the relative frequency of each canonical amino acid for each sequence in the current alignment. The Plotly library allows a user to select a subset of amino acids, download the Frequency box plot as PNG, show a description of the closest datapoint on hover, or display a comparison of amino acid frequencies and distribution statistics across all species.

### ProteoVision Data

*ProteoVision Inputs.* The DESIRE database provides a set of master alignments for every ribosomal protein. To pro-

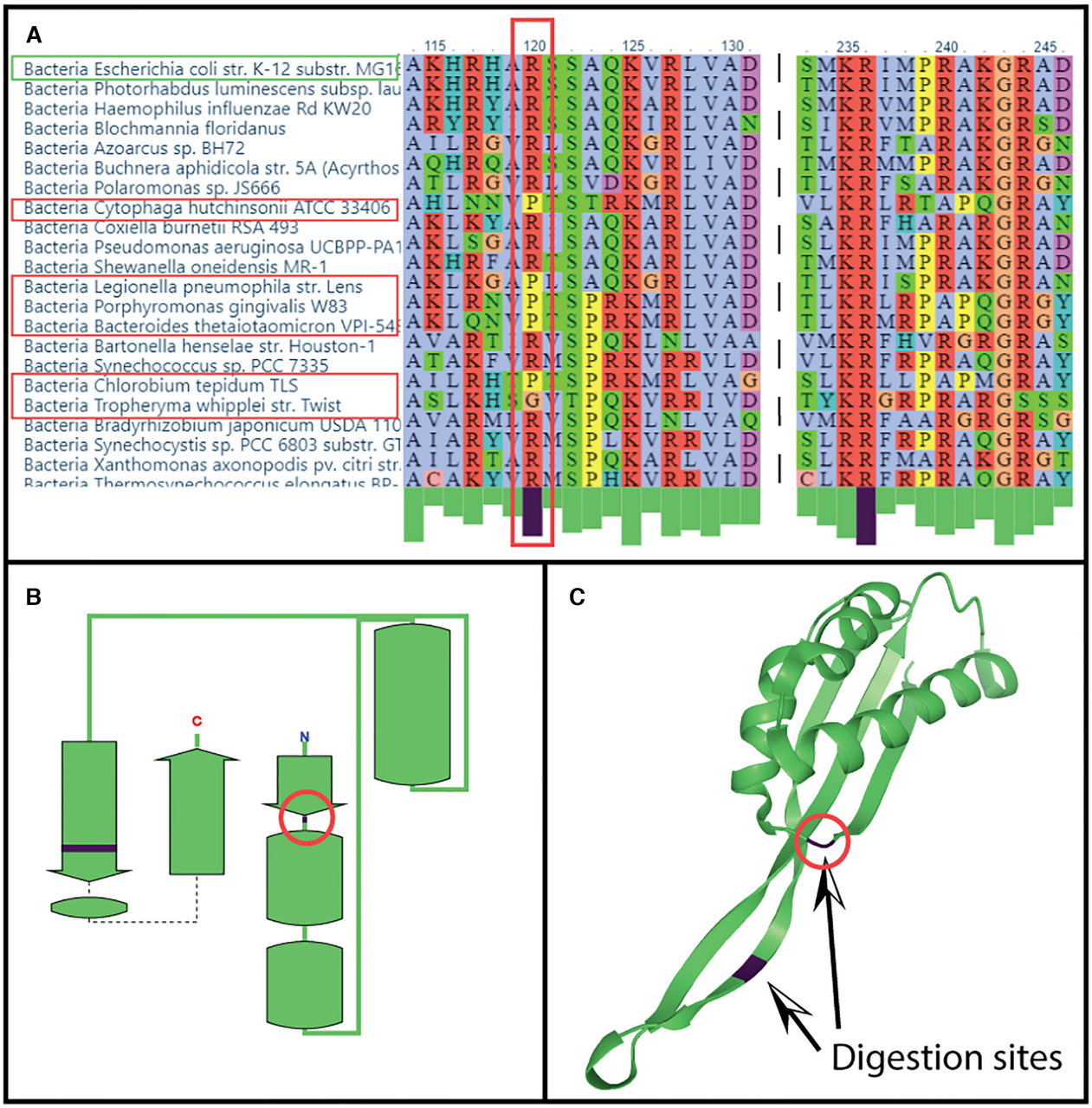



**Figure 3.** Visualization of custom data for a ribosomal protein. (**A**) DESIRE Alignment of rProtein uL22 from bacterial species, connected to (**B**) the topology diagram and (**C**) the 3D structure of *E. coli* uL22 from PDB entry 4V9D, chain CS. Proteolysis data (57) mapped with ProteoVision highlights the digestion sites with purple simultaneously at (A) primary, (B) secondary, and (C) tertiary levels of the uL22 structure. The digestion sites are indicated by arrows in panel (C) for clarity. The *E. coli* uL22 sequence is highlighted with green box in panel (A). Position 11 (*E. coli* numbering) is highlighted with red box in (A) and red circle in (B) and (C). Species with variable amino acid types at position 11 are also highlighted with red boxes in (A).

duce the master alignment, amino acid sequences of species from the DESIRE database were aligned using PROMALS3D (45) and curated with MATRAS (46). In User upload mode ProteoVision accepts alignments in FASTA format (up to 2000 sequences) and external 3D structures in PDB format (containing a single protein chain). In both modes ProteoVision supports the mapping of structure-indexed custom data, uploaded as a CSV. All custom data uploaded by the user (alignment or CSV) is removed from the server after the user ends their session.

*External APIs.* ProteoVision uses several external API resources. Identification of relevant PDB structures in User upload mode is done with the aid of the NCBI-BLAST+ API from EBI, and in DESIRE mode with help of the riboXYZ API. Topology and 3D structures are retrieved through an API service supported by EBI (27).

*Calculated data.* ProteoVision calculates gap-adjusted amino-acid frequencies from alignments and uses those frequencies to compute physicochemical properties and phylo-

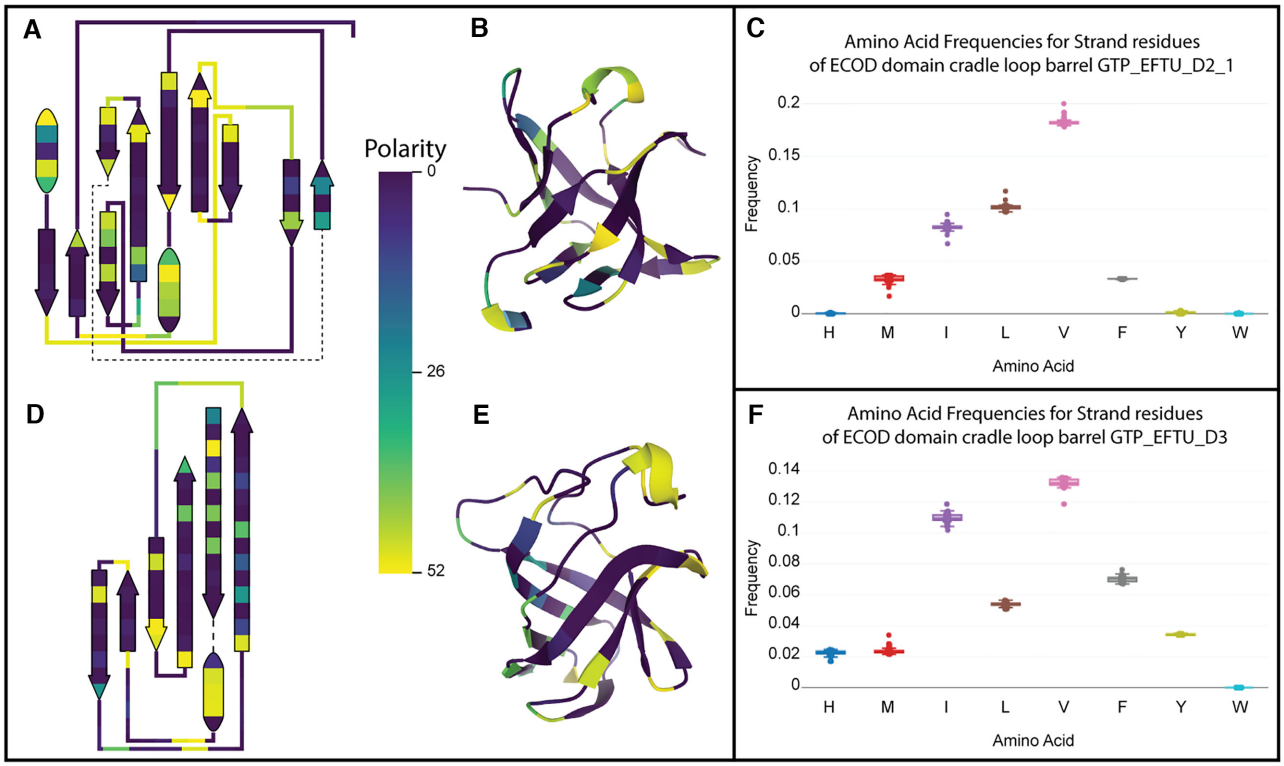



**Figure 4.** Simultaneous mapping of data from custom alignment of a non-ribosomal protein. Alignment of EF-Tu was used to select and visualize two cradle loop domains. *GTP_EFTU_D2_1;* (**A**) Topology diagram, (**B**) 3D Structural representation, and (**C**) Amino acid frequency distribution of aromatic and hydrophobic residues from β-strand-localized residues. *GTP_EFTU_D3;* (**D**) Topology diagram, (**E**) 3D structural representation and (**F**) amino acid frequency distribution. Amino acid polarity is represented by a color gradient from purple (low) to yellow (high).

genetic conservation scores. By default, ProteoVision uses the clustered sequences, but the user has an option to use their original alignment for visualization and calculation of properties. ProteoVision supports calculation of physicochemical properties including (47) charge, (48) hydropathy, (49) hydrophobicity, (50) polarity and (51) mutability. Conservation within the clustered alignment is calculated as Shannon Entropy (52). In user upload mode, alignment sequences are clustered with CD-Hit (53–55), using a 90% identity threshold. Conservation scores that explicitly take into account evolutionary relationships between the aligned sequences (e.g. ConSurf or Zebra2 (25,56)) can be computed externally, imported, mapped and visualized with User upload mode. The conservation/divergence signal between the two pre-defined groups is calculated with TwinCons.

*ProteoVision outputs.* ProteoVision provides the ability to save all computed and visualized data in multiple formats. In addition to saving images from every applet, the web server can generate a PyMOL script that fetches the selected PDB ID, extracts the selected chain, and applies the colors for each calculated property. All calculated data and amino acid frequencies are available for download as a CSV file and the current alignment can be downloaded in FASTA format. Finally, ProteoVision can output a session file in JSON format. The session file stores information about the current alignment, structure, and calculated properties. This allows the user to restore a state of work and progress from a previously saved session file.

## EXAMPLE APPLICATIONS

To demonstrate the utility of ProteoVision, we report three example applications. These examples illustrate the ability of ProteoVision to map custom data in the DESIRE mode and to link a custom alignment with a chosen 3D structure in the User upload mode. We provide external data files and session files for each example as supplementary datasets (Supp. Data S2–7).

### Example 1: visualization of custom data for a ribosomal protein

Trypsin is a protease that cleaves a protein chain at a lysine or arginine residue. Here, we visualize the experimental results of trypsin proteolysis for uL22 of *Escherichia coli*, which is digested at positions 11 and 84 (57). The alignment of rProtein uL22 for all bacterial species in DESIRE database and its structure from *E. coli* (PDB 4V9D) are selected for sequence, topology, and 3D visualizations. Digestion results are uploaded to ProteoVision as a CSV file (Supp. Data S4), in which values for the digestion site residues are set to zero, and values for the remaining residues are set to 0.75. ProteoVision associates these data with a color palette and draws the color-encoded data on the bar-plot of MSAViewer as well as the structures within

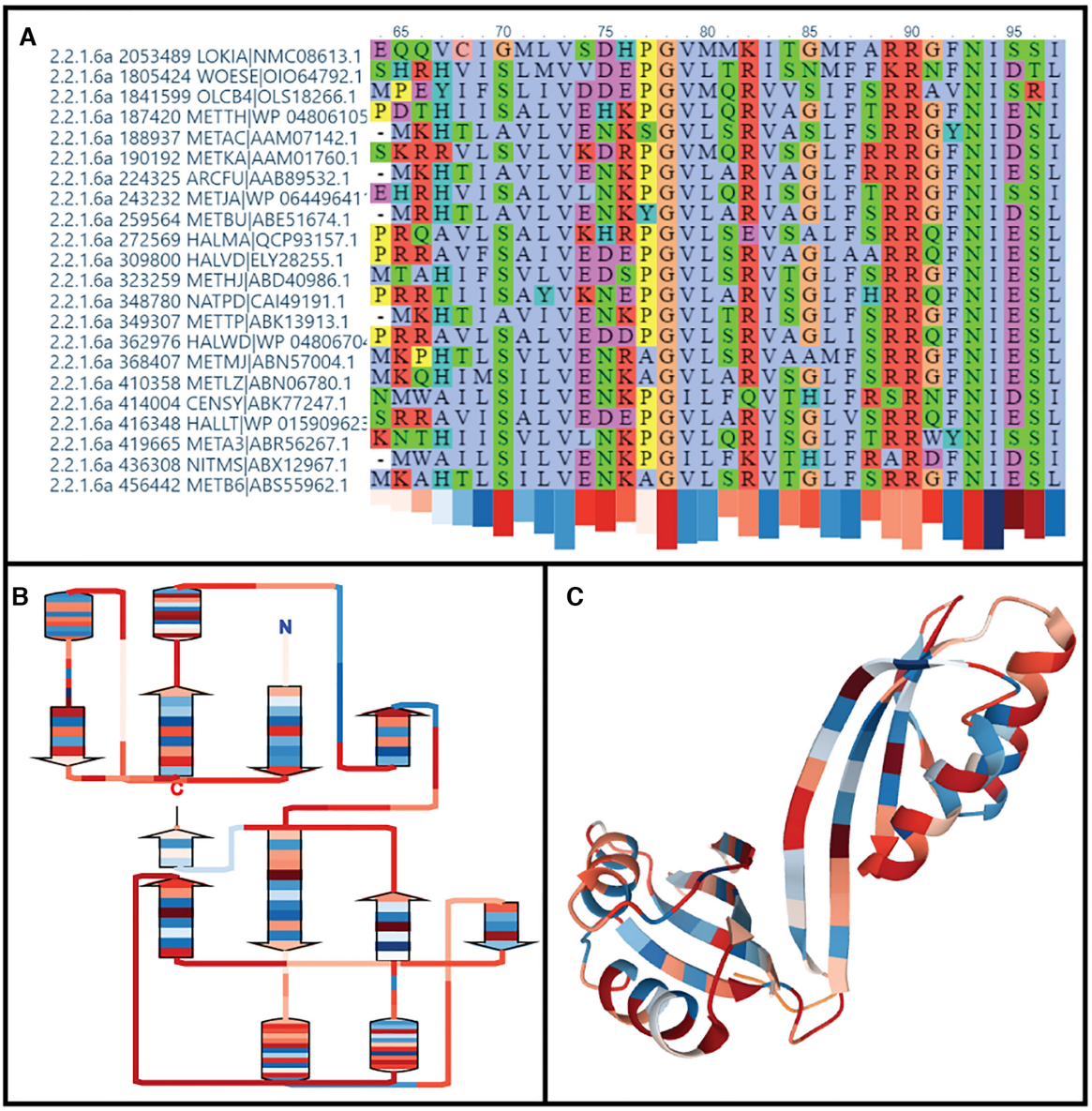



**Figure 5.** Visualization of custom alignment and structure in User upload mode. (**A**) Alignment of acetolactate synthase from 25 archaeal species. (**B**) Automatically generated topology diagram using Pro-Origami from custom structure. (**C**) Custom 3D structure modelled with Swiss-Model from a sequence of acetolactate synthase (NMC08613.1) from Candidatus Lokiarchaeota.

Topology and 3D viewers (Figure. 3 A–C). ProteoVision confirms that uL22 digestion occurs at the arginine residues 11 and 84. The 3D representation demonstrates that digestion sites are not buried within the uL22 core.

### Example 2: visualization of a non-ribosomal protein from a custom alignment

EF-Tu is an auxiliary protein of the translational system, comprised of two cradle loop β-barrel domains and one P-loop domain (58). Here, we compare polarity and amino acid frequencies for the pair of cradle loop β-barrel domains *GTP_EFTU_D2_1* and *GTP_EFTU_D3* (28). Upon uploading a custom alignment (Supp. Data S5) of EF-Tu, which contains 67 bacterial sequences, ProteoVision identifies relevant 3D structures with NCBI-BLAST. Using EF-Tu structure PDB ID 1EFC (59), ProteoVision calculates the alignment-structure mapping, as well as conservation and physicochemical properties. The web server further identifies ranges for three ECOD domains in the selected structure, truncates the structure by the range of a selected domain, and maps the polarity onto structural representations. The resulting visualizations reveal that polarity is correlated with protein secondary and tertiary structure: the core of the β-barrel domains is comprised of beta strands and they are predominantly nonpolar (Figure 4A and D), while the solvent exposed loop and helix regions tend to contain more polar residues (Figure 4B and E). ProteoVi-

sion calculates amino acid frequency distributions based on secondary structure elements in a chosen domain (Figure 4C and F).

**Example 3: visualization of custom alignment and structure**

Acetolactate synthase (EC 2.2.1.6) is an enzyme that catalyzes the first step of the branched amino acid synthetic pathway. To demonstrate ProteoVision's utility in visualizing modelled custom structures we generated a 3D model with Swiss-Model (60) from a sequence of acetolactate synthase (NMC08613.1) from *Candidatus* Lokiarchaeota (TAX ID: 2053489). Upon uploading a custom alignment for 24 archaeal sequences of acetolactate synthase (Supp Data S6), and the modeled 3D structure (Supp. Data S7), ProteoVision generates a topology diagram and visualizes the alignment, topology diagram and 3D structure in their respective viewers; ProteoVision connects equivalent residue positions within the viewers. Following the main pipeline, ProteoVision provides an option to compute available physico-chemical properties or to map user upload data. CD-HIT analysis reveals 22 cluster groups that are used to compute hydrophobicity. A fragment of the alignment, topology diagram and 3D structure of acetolactate synthase from *Candidatus* Lokiarchaeota with mapped hydrophobicity values are shown in Figure 5A–C.

## DISCUSSION

Here we present ProteoVision – a web server designed for visualization of rProteins. ProteoVision enables users to explore rProteins across phylogeny and to visualize related information at levels of primary, secondary, and tertiary structure. Additionally, ProteoVision supports custom alignments and data for any protein, making it a general tool for protein visualization and data mapping. ProteoVision shares visualization philosophy with the previously established RiboVision web server (26) and complements it in terms of the ribosomal data offered to the community. ProteoVision provides an API service for rProtein nomenclature, sequences, alignments, and annotations. The API is available at https://proteovision.chemistry.gatech.edu/desire-api/.

Molecular visualization is a quickly developing area of biological research. Advances of JavaScript and WebGL technologies have given rise to numerous platforms for visualization of macromolecules. Among them are alignment viewers (JalView (61,62), MSAViewer (30)), topology diagram viewers (PDBeSum (31,63), ProOrigami (41)) and 3D viewers (NGL (64), LiteMol (65,66)). Some of these tools have been connected to web servers (PDBe (27), Zebra2 (56), Consurf (25), Aquaria (67)) that enable users to visualize a given structure online in multiple representations. We have extended some of the available functionalities by integrating existing visualization applets into a single server, which provides the ability to explore structural features across the tree of life, to map phylogenetic or physicochemical properties or custom data, and to save the resulting visualizations in a variety of formats as publication quality images. We hope that ProteoVision will meet the needs of structural and evolutionary biologists (including those who work in the origins of life field) and may also be of interest to a broad audience that enjoys molecular visualizations.

## DATA AVAILABILITY

The web server is available at https://proteovision.chemistry.gatech.edu. The project development repository is available at https://github.com/LDWLab/DESIRE. ProteoVision is free and open to all users and there is no login required. Supplementary Information and datasets are available at NAR Online.

## NOTE ADDED IN PROOF

Just before submitting the proofs of the current manuscript, we noticed that publication on Mol* Viewer appeared on line in same web server issue of NAR. Thus, in addition to ref. 32, Mol* should be referenced as:

Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, Alexander S Rose. Nucleic Acids Research, gkab314, https://doi.org/10.1093/nar/gkab314.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We would like to thank Prof. Khanh Dao Duc, Artem Kushner, Mr. Rohan Gupta, Joshua Lin, the group of Prof. George E. Fox, and Yulia Dumov, whose input has been invaluable to our work.

*Author contributions*: P.I.P. and C.A.-C. compiled protein alignments; P.I.P., C.R.B., V.L.C., M.A., A.S., C.A.-C. and A.S.P. developed DESIRE database; P.I.P. and C.D.M. developed the back end; P.I.P., A.S.P., H.M.M., C.D.M., A.M. and B.G. adjusted viewers and integrated the front end; P.I.P., H.M.M., C.D.M., A.M. and C.R.B. wrote and compiled the documentation; P.I.P., H.M.M., C.D.M., C.A.-C., L.D.W., A.S.P. wrote the manuscript; L.D.W. and A.S.P. oversaw the project and acquired the funding.

## FUNDING

National Aeronautics and Space Administration [80NSSC18K1139 to L.D.W., A.S.P.]; C.A.-C. was supported by the NASA Postdoctoral Program, administered by Universities Space Research Association under contract with NASA. Funding for open access charge: National Aeronautics and Space Administration [80NSSC18K1139].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Pauling,L. and Corey,R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl Acad. Sci. U.S.A.*, **37**, 251–256.

2. Pauling,L., Corey,R.B. and Branson,H.R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl Acad. Sci. U.S.A.*, **37**, 205–211.
3. Kendrew,J.C., Bodo,G., Dintzis,H.M., Parrish,R., Wyckoff,H. and Phillips,D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181**, 662–666.
4. Green,D.W., Ingram,V.M., Perutz,M.F. and Bragg,W.L. (1954) The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. *Proc. R. Soc. London Ser. A*, **225**, 287–307.
5. Murata,M., Richardson,J.S. and Sussman,J.L. (1985) Simultaneous comparison of three protein sequences. *Proc. Natl Acad. Sci. U.S.A.*, **82**, 3073–3077.
6. Richardson,J.S. (1977) B-Sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.
7. Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.
8. Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F. and Yonath,A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679–688.
9. Agmon,I., Bashan,A. and Yonath,A. (2006) On ribosome conservation and evolution. *Isr. J. Ecol. Evol.*, **52**, 359–374.
10. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
11. Wimberly,B.T., Brodersen,D.E., Clemons,W.M., Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30s ribosomal subunit. *Nature*, **407**, 327–339.
12. Frank,J., Zhu,J., Penczek,P., Li,Y., Srivastava,S., Verschoor,A., Radermacher,M., Grassucci,R., Lata,R.K. and Agrawal,R.K. (1995) A model of protein synthesis based on cryo-electron microscopy of the *E. coli* ribosome. *Nature*, **376**, 441–444.
13. Moore,P.B. (2012) How should we think about the ribosome? *Annu. Rev. Biophys.*, **41**, 1–19.
14. Woese,C.R. and Fox,G.E. (1977) The concept of cellular evolution. *J. Mol. Evol.*, **10**, 1–6.
15. Ben-Shem,A., Garreau de Loubresse,N., Melnikov,S., Jenner,L., Yusupova,G. and Yusupov,M. (2011) The structure of the Eukaryotic ribosome at 3.0 Å resolution. *Science*, **334**, 1524–1529.
16. Klein,D.J., Moore,P.B. and Steitz,T.A. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.*, **340**, 141–177.
17. Lecompte,O., Ripp,R., Thierry,J.C., Moras,D. and Poch,O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.*, **30**, 5382–5390.
18. Vishwanath,P., Favaretto,P., Hartman,H., Mohr,S.C. and Smith,T.F. (2004) Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. *Mol. Phylogen. Evol.*, **33**, 615–625.
19. Yutin,N., Puigbò,P., Koonin,E.V. and Wolf,Y.I. (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, **7**, e36972.
20. Melnikov,S., Manakongtreecheep,K. and Söll,D. (2018) Revising the structural diversity of ribosomal proteins across the three domains of life. *Mol. Biol. Evol.*, **35**, 1588–1598.
21. Roberts,E., Sethi,A., Montoya,J., Woese,C.R. and Luthey-Schulten,Z. (2008) Molecular signatures of ribosomal evolution. *Proc. Natl Acad. Sci. U.S.A.*, **105**, 13953–13958.
22. Timsit,Y., Sergeant-Perthuis,G. and Bennequin,D. (2021) Evolution of ribosomal protein network architectures. *Sci. Rep.*, **11**, 625.
23. Bowman,J.C., Petrov,A.S., Frenkel-Pinter,M., Penev,P.I. and Williams,L.D. (2020) Root of the tree: the significance, evolution, and origins of the ribosome. *Chem. Rev.*, **120**, 4848–4878.
24. Martinez,X., Chavent,M. and Baaden,M. (2020) Visualizing protein structures — tools and trends. *Biochem. Soc. Trans.*, **48**, 499–506.
25. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344–W350.
26. Bernier,C.R., Petrov,A.S., Waterbury,C.C., Jett,J., Li,F., Freil,L.E., Xiong,X., Wang,L., Migliozzi,B.L.R., Hershkovits,E. *et al.* (2014) Ribovision suite for visualization and analysis of ribosomes. *Faraday Discuss.*, **169**, 195–207.
27. Mir,S., Alhroub,Y., Anyango,S., Armstrong,D.R., Berrisford,J.M., Clark,A.R., Conroy,M.J., Dana,J.M., Deshpande,M., Gupta,D. *et al.* (2017) PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.*, **46**, D486–D492.
28. Cheng,H., Schaeffer,R.D., Liao,Y., Kinch,L.N., Pei,J., Shi,S., Kim,B.-H. and Grishin,N.V. (2014) Ecod: an evolutionary classification of protein domains. *PLoS Comp Biol*, **10**, e1003926.
29. Kovacs,N.A., Petrov,A.S., Lanier,K.A. and Williams,L.D. (2017) Frozen in time: the history of proteins. *Mol. Biol. Evol.*, **34**, 1252–1260.
30. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
31. Laskowski,R.A., Jabłońska,J., Pravda,L., Vařeková,R.S. and Thornton,J.M. (2018) Pdbsum: structural summaries of Pdb entries. *Protein Sci.*, **27**, 129–134.
32. Sehnal,D., Rose,A.S., Koča,J., Burley,S.K. and Velankar,S. (2018) In: *Proc Wkshp Mol Graph Vis Anal Mol Graph*. Eurographics Association, Brno, Czech Republic, pp. 29–33.
33. Cheng,H., Liao,Y., Schaeffer,R.D. and Grishin,N.V. (2015) Manual classification strategies in the Ecod database. *Proteins: Struct. Funct. Bioinform.*, **83**, 1238–1251.
34. Bernier,C.R., Petrov,A.S., Kovacs,N.A., Penev,P.I. and Williams,L.D. (2018) Translation: the universal structural core of life. *Mol. Biol. Evol.*, **35**, 2065–2076.
35. Spang,A., Saw,J.H., Jørgensen,S.L., Zaremba-Niedzwiedzka,K., Martijn,J., Lind,A.E., van Eijk,R., Schleper,C., Guy,L. and Ettema,T.J.G. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.
36. Penev,P.I., Fakhretaha-Aval,S., Patel,V.J., Cannone,J.J., Gutell,R.R., Petrov,A.S., Williams,L.D. and Glass,J.B. (2020) Supersized ribosomal RNA expansion segments in Asgard archaea. *Genome Biol Evol*, **12**, 1694–1710.
37. Burki,F., Roger,A.J., Brown,M.W. and Simpson,A.G.B. (2020) The new tree of eukaryotes. *Trends Ecol. Evol.*, **35**, 43–55.
38. Ban,N., Beckmann,R., Cate,J.H.D., Dinman,J.D., Dragon,F., Ellis,S.R., Lafontaine,D.L.J., Lindahl,L., Liljas,A., Lipton,J.M. *et al.* (2014) A new system for naming ribosomal proteins. *Curr. Opin. Struct. Biol.*, **24**, 165–169.
39. Federhen,S. (2011) The NCNI Taxonomy Database. *Nucleic Acids Res.*, **40**, D136–D143.
40. Kovacs,N.A., Penev,P.I., Venapally,A., Petrov,A.S. and Williams,L.D. (2018) Circular permutation obscures universality of a ribosomal protein. *J. Mol. Evol.*, **86**, 581–592.
41. Stivala,A., Wybrow,M., Wirth,A., Whisstock,J.C. and Stuckey,P.J. (2011) Automatic generation of protein structure cartoons with pro-origami. *Bioinformatics*, **27**, 3315–3316.
42. Dwyer,T., Marriott,K. and Wybrow,M. (2009) Dunnart: A Constraint-Based Network Diagram AuthoringTool. In: Tollis,I.G. and Patrignani,M. (eds). *Graph Drawing. GD 2008. Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, Vol. **5417**, pp. 420–431.
43. Katoh,K. and Standley,D.M. (2013) Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
44. Velankar,S., van Ginkel,G., Alhroub,Y., Battle,G.M., Berrisford,J.M., Conroy,M.J., Dana,J.M., Gore,S.P., Gutmanas,A., Haslam,P. *et al.* (2015) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.
45. Pei,J. and Grishin,N.V. (2014) In: Russell,D.J. (ed). *Multiple Sequence Alignment Methods*. Humana Press, Totowa, NJ, pp. 263–271.
46. Kawabata,T. (2003) Matras: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.
47. Sehnal,D., Svobodová Vařeková,R., Berka,K., Pravda,L., Navrátilová,V., Banáš,P., Ionescu,C.-M., Otyepka,M. and Koča,J. (2013) Mole 2.0: advanced approach for analysis of biomacromolecular channels. *J. Cheminformatics*, **5**, 39.
48. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
49. Cid,H., Bunster,M., Canales,M. and Gazitúa,F. (1992) Hydrophobicity and structural classes in proteins. *Protein Eng. Des. Select.*, **5**, 373–375.

50. Zimmerman,J.M., Eliezer,N. and Simha,R. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.
51. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, **8**, 275–282.
52. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
53. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
54. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
55. Li,W. and Godzik,A. (2006) CD-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
56. Suplatov,D., Sharapova,Y., Geraseva,E. and Švedas,V. (2020) Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies. *Nucleic Acids Res.*, **48**, W65–W71.
57. Hamburg,D.M., Suh,M.J. and Limbach,P.A. (2009) Limited proteolysis analysis of the ribosome is affected by subunit association. *Biopolymers*, **91**, 410–422.
58. Weijland,A., Harmark,K., Cool,R.H., Anborgh,P.H. and Parmeggiani,A. (1992) Elongation factor Tu: a molecular switch in protein biosynthesis. *Mol. Microbiol.*, **6**, 683–688.
59. Song,H., Parsons,M.R., Rowsell,S., Leonard,G. and Phillips,S.E.V. (1999) Crystal structure of intact elongation factor EF-Tu from *Escherichia coli* in GDP conformation at 2.05Å resolution. *J. Mol. Biol.*, **285**, 1245–1256.
60. Waterhouse,A., Bertoni,M., Bienert,S., Studer,G., Tauriello,G., Gumienny,R., Heer,F.T., de Beer,T.A.P., Rempfer,C., Bordoli,L., Lepore,R. and Schwede,T. (2018) Swiss-Model: homology modelling of protein structures and complexes. *Nucleic Acids Res.*, **46**, W296–W303.
61. Procter,J.B., Carstairs,G.M., Soares,B., Mourão,K., Ofoegbu,T.C., Barton,D., Lui,L., Menard,A., Sherstnev,N., Roldan-Martinez,D. *et al.* (2021) Alignment of biological sequences with Jalview. *Methods Mol. Biol.*, **2231**, 203–224.
62. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
63. Laskowski,R.A. (2008) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
64. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlić,A. and Rose,P.W. (2018) NGL Viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
65. Sehnal,D., Deshpande,M., Vařeková,R.S., Mir,S., Berka,K., Midlik,A., Pravda,L., Velankar,S. and Koča,J. (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods*, **14**, 1121–1122.
66. Sehnal,D., Svobodová,R., Berka,K., Pravda,L., Midlik,A. and Koča,J. (2020) In: Gáspári,Z. (ed). *Structural Bioinformatics: Methods and Protocols*. Springer US, NY, pp. 1–13.
67. O'Donoghue,S.I., Sabir,K.S., Kalemanov,M., Stolte,C., Wellmann,B., Ho,V., Roos,M., Perdigão,N., Buske,F.A., Heinrich,J. *et al.* (2015) Aquaria: simplifying discovery and insight from protein structures. *Nat. Methods*, **12**, 98–99.