

# RiboXYZ: a comprehensive database for visualizing and analyzing ribosome structures

Artem Kushner<sup>1,2</sup>, Anton S. Petrov<sup>3,4,5</sup> and Khanh Dao Duc<sup>1,2,\*</sup>

<sup>1</sup>Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada, <sup>2</sup>Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, <sup>3</sup>NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA, <sup>4</sup>School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA and <sup>5</sup>School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received August 23, 2022; Revised September 19, 2022; Editorial Decision September 30, 2022; Accepted October 12, 2022

## ABSTRACT

Recent advances in Cryo-EM led to a surge of ribosome structures deposited over the past years, including structures from different species, conformational states, or bound with different ligands. Yet, multiple conflicts of nomenclature make the identification and comparison of structures and ortholog components challenging. We present RiboXYZ (available at <https://ribosome.xyz>), a database that provides organized access to ribosome structures, with several tools for visualisation and study. The database is up-to-date with the Protein Data Bank (PDB) but provides a standardized nomenclature that allows for searching and comparing ribosomal components (proteins, RNA, ligands) across all the available structures. In addition to structured and simplified access to the data, the application has several specialized visualization tools, including the identification and prediction of ligand binding sites, and 3D superimposition of ribosomal components. Overall, RiboXYZ provides a useful toolkit that complements the PDB database, by implementing the current conventions and providing a set of auxiliary tools that have been developed explicitly for analyzing ribosome structures. This toolkit can be easily accessed by both experts and non-experts in structural biology so that they can search, visualize and compare structures, with various potential applications in molecular biology, evolution, and biochemistry.

## INTRODUCTION

The ribosome is a universal RNA-protein complex, whose structure has been a major focus of study for the past decades (1). Due to the recent progress in imaging technologies and the notable renaissance of cryogenic elec-

tron microscopy (cryo-EM) (2), more than a thousand of high-resolution ribosome structures are currently deposited in the protein data bank (PDB) (3), encompassing various species, as well as various binding and conformational states (4).

While this variety provides an exciting opportunity to examine specific components of the ribosome through the lens of any single structure available, integrating the data from the PDB for such analysis can be challenging in practice. It is even more challenging to develop computational pipelines that would provide a comparison across multiple structures selected by a user. More precisely, identifying the different components—e.g. rRNA, proteins or ligands—of a given structure can require some manual curation, due to different ontologies of its components. While a ribosome structure-based nomenclature has been proposed (5), the PDB still contains numerous references to earlier naming systems, which lead to unrelated ribosomal proteins being assigned the same name across deposited structures. Ambiguities and inconsistencies are similarly present in the most standard protein databases that include ribosome protein families, such as Pfam (6) or UniProt (7), and have explicitly been mentioned as an obstacle for the ribosome structure-based nomenclature system to be adopted (8).

To facilitate this access and visualization of ribosome components over the whole set of available structures, we present here RiboXYZ (publicly available at <https://ribosome.xyz>), a database that provides organized access to ribosome structures and enables a direct search and comparison of these structures and their components. In addition, RiboXYZ proposes several export facilities and visualization tools, including the identification and prediction of ligand binding sites, and 3D superimposition of structures and ribosomal components. As illustrated with various examples and applications below, RiboXYZ provides a useful toolkit that complements the PDB database by implementing the current conventions and providing a set of auxiliary tools that have been developed explicitly for ribo-

\*To whom correspondence should be addressed. Tel: +1 604 822 0683; Email: [kdd@math.ubc.ca](mailto:kdd@math.ubc.ca)

somal structures. This toolkit can be easily accessed by both experts and non-experts in structural biology so that they can search, visualize and compare structures, with various potential applications in molecular biology, evolution, and biochemistry.

## WEB SERVER AND DATABASE DESCRIPTION

RiboXYZ is a database application that provides organized access to ribosome structures, with several tools for visualisation and export. The user can interact with the database from the application website. The website is mainly divided into two sections, *Database* and *Tools*, whose pages are accessible from the main navigation panel. The database section provides separate access to whole structures, RNA's, and proteins. The *Tools* section allows to visualize and compare the elements of the database and is composed of three modules which all integrate a 3D viewer (9): *Visualization*, *3D superimposition* and *Ligands/Binding Sites*. All these pages are accessible from the RiboXYZ's home page after opening a left tab menu. The menu also contains a link to RiboXYZ's *How to* page, that contains various tutorials and a user manual (see Supplementary File 1) providing a complete description of the application and its functionalities, as summarized next.

### General infrastructure and development

RiboXYZ is composed of three services, as shown in Figure 1A: the website itself, the Django server and the Neo4j graph-database. The user-facing website is a single-page application (SPA) developed using the React framework for presentational components in front end, and the Redux library for state-management. The Django web-server provides structural computation and exposes the API endpoints. This web-server has access to both the database and static files, with Python libraries (NumPy, SciPy, Biopython) used for most of the computation. The database used is Neo4j and its schema definition language is Cypher, a graph-oriented variant of SQL. Ribosome structures were initially obtained from RCSB PDB's Web Service (10), and processed so the structures and their sub-components are nodes of the semantic graph database (as described next in more details). In addition, the database uses Pfam entries (6) to identify ribosomal proteins as well as select RCSB annotations imported from other resources like Uniprot (7) and Gene Ontology (11). The application is deployed to a *c5a.xlarge* instance with 8 GB of memory and 120GB of disk space storage of the AWS public cloud, which we intend to scale up after initial release.

### Data structure

Ribosome structures were initially obtained as *.cif* files from the PDB. For each PDB entry, a semantic *profile* of the structure was created and uploaded to the database, that includes annotations provided by the authors. Additionally, external resources associated with the specific PDB entry were obtained as a *.json* file. Individual chains present in each structure are integrated to the database and sorted into various classes, according to the diagram shown in Figure 1B. Thus, aside from whole structures, the semantic

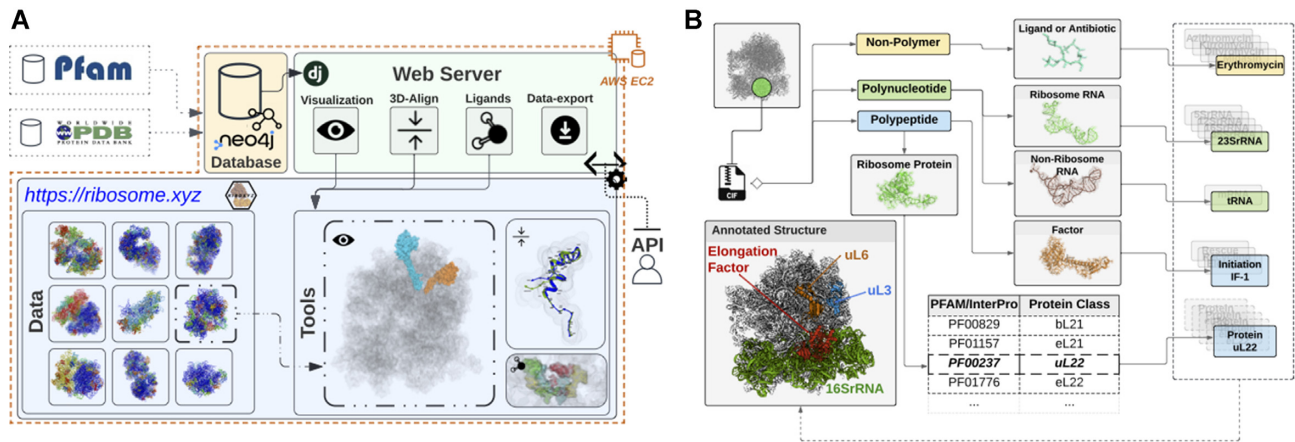
database includes individual polymer chains, such as proteins and RNA, as well as non-polymer entities (ligands, ions, antibiotics and other small molecules). In particular, more specific attention was given to classifying protein chains, due to the same class of ribosomal protein carrying different names in different structures (for example, ribosomal protein eL21 provided as *60S ribosomal protein L21-A*, *60S ribosomal protein L21* or *KLLA0E23651p* in structures PDB 4U4Y, 6LQM and 5IT7 respectively) and, conversely, several instances of ribosomal proteins from different species carrying similar names despite being unrelated in structure and function (like *Ribosomal protein L10 (Predicted)* and *50S ribosomal protein L10* in eukaryotic structure 6R7Q and bacterial structure 5AFI). Although a universal naming system was proposed by a consortium of structural biologists and biochemists (5), applying this consensus retroactively at scale to data already deposited to PDB and elsewhere has seemed prohibitive so far. Thus, our database implements a reduced version of this universal nomenclature upon re-annotating structures from the PDB, by cross-matching a protein's PFAM families against a manually curated table that maps PFAM clusters (6) to the proposed proteins classes, as illustrated in Figure 1B.

### Database content

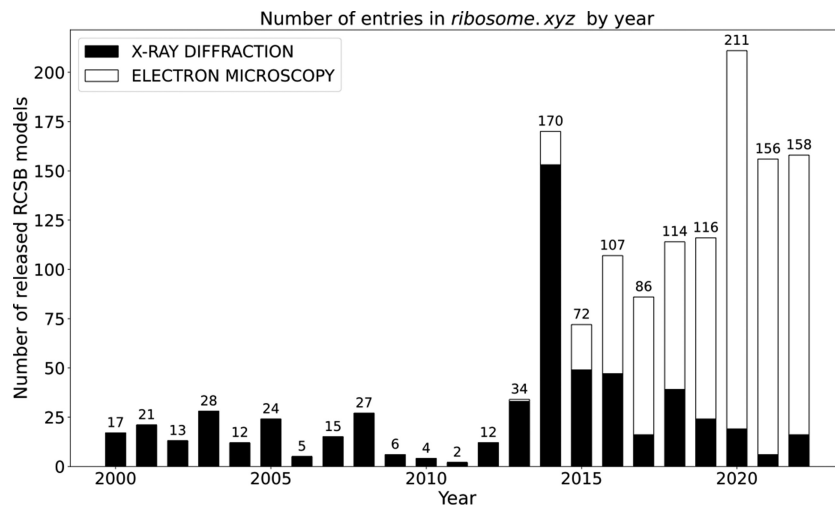
As of August 2022, the database contains >1100 structures with a resolution that is less than 4 Å, that yield ~65 000 single protein chains and 5300 RNA chains. Note that the identification of the chains might be missing in most recent structures, due to temporarily incomplete labeling of polymer entities with Pfam entries in the GraphQL endpoint of the RCSB's data API (10). Structures were selected from depositions made from 2000 (with the highest resolution obtained so far at 1.5 Å for X-ray structures and 1.98 Å for Cryo-EM). As shown in Figure 2, the number of ribosomal PDB entries has significantly increased over the past decade, with half of the structures of the database having been deposited in the past three years, and more than 90 % from Cryo-EM. In Table 1, we also detail how these structures are distributed among different species, with more than 59% of structures accounting for *E. coli*, *T. thermophilus*, *S. cerevisiae* and *H. sapiens* as the four major species studied. We anticipate our database to steadily grow and will perform monthly updates to import new depositions from the PDB, via a job scheduler on the server. Also note that as we are preparing the current submission, the Pfam website just announced to be decommissioned, which as a result will lead our future updates to rely on Interpro (12), as the source of future Pfam annotations.

### Requirements and files format

The current version of the website has been developed for and tested in Firefox 104.0 (64-bit) web-browser. We recommend to use the website with Mozilla Firefox. Other browsers (e.g. Safari) might have formatting differences. Multiple file-formats are available for download, with open source softwares available to open and further process them: *.fasta* files are used for representing nucleotide or amino acid (protein) sequences. The re-annotated structures from



**Figure 1.** Architecture and data processing of RiboXYZ. (A) Diagram that describes RiboXYZ's architecture. We process data from the PDB and Pfam entries into a graph database structure that is composed of edited structures and their components, which are analyzed for different applications using the web server. The results are mainly accessible to the user via a web interface that give access to both data and visualization tools. (B) File processing. Structural files (.cif) are processed to first classify chain units. Upon identifying these components, we further edit the chain ID's of the original structure with our common nomenclature, to facilitate comparison across structures.



**Figure 2.** Number of ribosomal structures in the database deposited in the PDB over the past 20 years.

**Table 1.** Structures in the database sorted by species/domain of life

Species	Domain	Total Count	X-ray	EM	Average resolution (X-ray/EM) in Å
<i>E. coli</i> (K-12)	Bacteria	265	45	220	3.2/3.3
<i>T. thermophilus</i>	Bacteria	230	221	9	3.1/3.6
<i>H. sapiens</i>	Eukarya	103	0	103	NA/3.2
<i>S. cerevisiae</i>	Eukarya	81	33	48	3.1/3.5
<i>H. marismortui</i>	Archaea	65	65	0	2.8/NA
Others	All	393	76	317	3.2/3.2

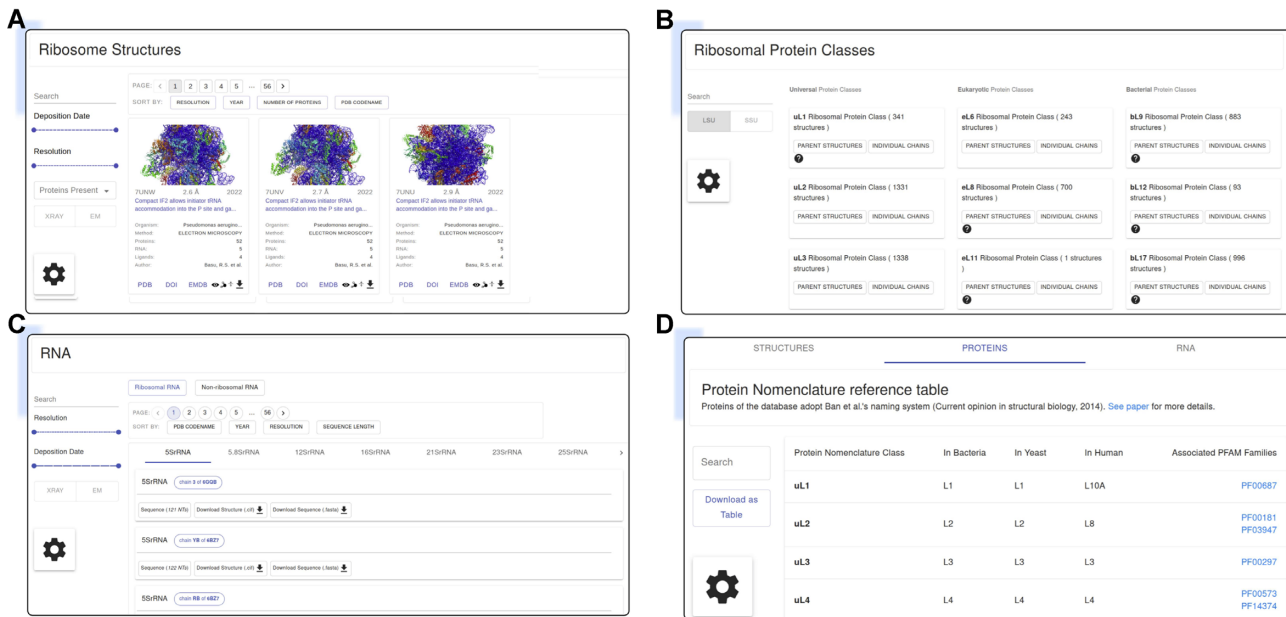
the database are in the .cif format. Image export from Mol\* viewer is in .png, and tables are finally provided in .csv format.

## GENERAL FEATURES

### Database access and search

The database section is mainly composed of four pages, shown in Figure 3A–D, that provide separate access to

whole structures, RNA's, and proteins, as well as a detailed description of the protein nomenclature system used to identify the different chains. Each page can be accessed from the website drawer menu, with a set of filters and a search tool bar available to run a search over the database (e.g. by typing species of interest). The pages also contain more detailed information on the structures and their components, and have links to export the data. To facilitate the identification of all the components, the *Nomenclature* page,



**Figure 3.** Data section of RiboXYZ, composed of: (A) the structure page that provides access to updated ribosome structures, (B) the protein page that covers all the different ribosome proteins, with access to parent structures and individual chains, (C) the RNA page that covers all the different RNA components present in the structures of the database (including both ribosomal and non-ribosomal RNA), and (D) the nomenclature page that displays the naming system used for ribosome proteins (see also (5) and the *Data structure* section) and RNA's, and provides for each structure the corresponding mapping with the chain ID's from the original PDB file.

shown in Figure 3(D), displays the naming system used for ribosome proteins and RNA's (5). In addition, this page also provides for each structure the mapping between the protein chain structures ID from the original PDB file and the corresponding protein in the nomenclature.

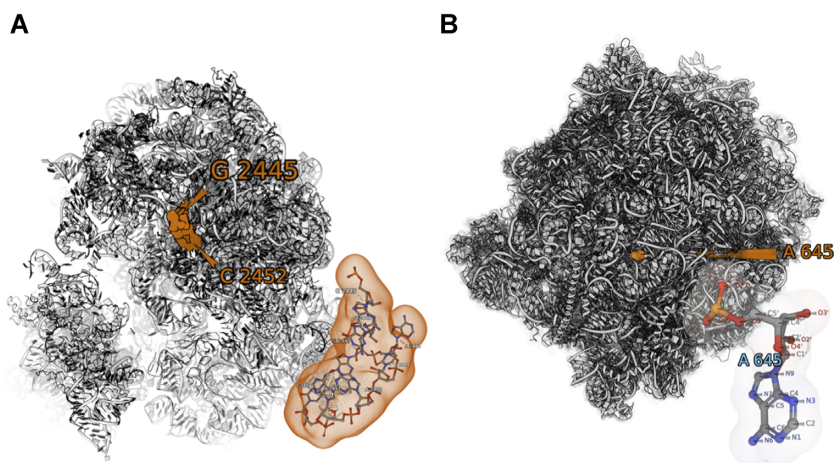
Each of the structure, protein and RNA pages has specific features. Each structure is represented by an *ID card* containing deposition metadata (e.g. organism, resolution, experimental method, publication info), and links to external resources (PDB, doi, EMBD). Clicking on each ID card displayed on the main screen opens a structure-specific page that lists the structure components. Structures can be downloaded as .cif files with chain names edited to match with the standard nomenclature. Each class of ribosomal proteins is also represented by a card, which contains the number of structures in which proteins of this class are present and Pfam annotation. From each class, the user can get access to the parent structures or individual chains. The RNA page has a specific arrangement, separating between ribosomal and non-ribosomal RNA's, with sub-tabs that correspond with classes (e.g. 5S rRNA, 5.8S rRNA etc. for ribosomal RNA, and mRNA/tRNA for non-ribosomal RNA). Each class links to a list of individual chains associated with the structures present in the database.

### Visualization tools

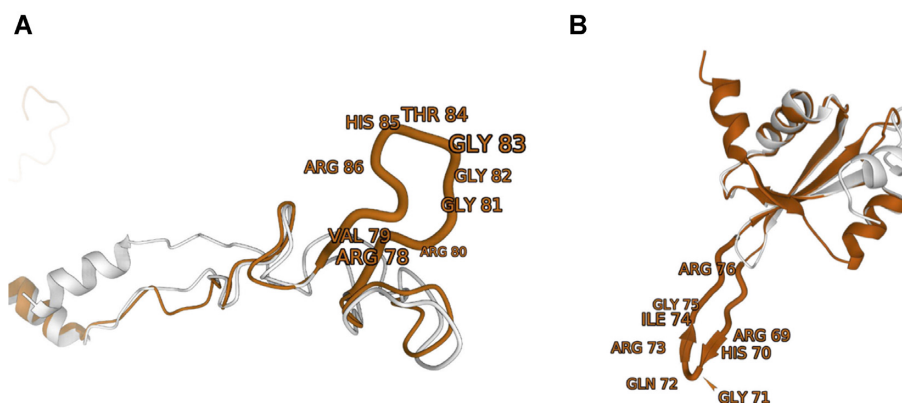
The *Tools* section allows to visualize and compare the elements of the database. This section contains three separate modules, named *Visualization*, *3D superimposition* and *Ligands/Binding Sites*, which can be accessed from the main navigation panel or the database section. We also illustrate each tool in Figures 4–6, in the context of some specific ex-

amples and applications detailed further below. The tools use the Mol\* viewer (9), with multiple features and options to represent the structure (e.g. mapping electrostatic surface potential, solvent accessibility etc.), as well as to annotate and export image. We refer to its documentation for more details. As detailed next, the *Superimposition* and *Ligands/Binding Sites* tools also perform sequence alignment between two given chains, with results that can directly be exported as a .csv file.

Each tool has its own specific features and purpose. The *Visualization* tool allows to select and visualize every single structures, and chain unit (e.g. protein, RNA, ligand) from the database. To do so, a searching tool is also available on the left menu. Once a structure is selected, the user can also highlight any chain, with a further option to select a specific region. Note that while these visualizations can also be done from the RCSB PDB, our database facilitates the search of adequate structure and identification of the sites, by using the same naming nomenclature for the different chain units. The *3D Superimpose* tool executes the 3D superimposition function *super* from an open source version of PyMOL, to spatially align two chains from two structures of the database. It is similar to the pairwise superimposition tool that was recently added to the PDB, but with a simplified interface and easier access to ribosomal components. The *Ligands/Binding Sites* tool is dedicated to the visualization of components which are neither RNA nor ribosomal protein strands, such as antibiotics, transfer and messenger RNA strands and elongation/initiation factors. These components can be selected and visualized within their parent structures. The tool also allows to predict the potential location of a ligand in any other target structure in which it does not exist. To do so, a binding site is first computed as



**Figure 4.** Examples of applications of the *Visualization* tool. (A) Localization of the ribosome PTC from PDB 4V7S by selecting nucleotides 2445–2452 from the 23S rRNA (see also Supplementary file 2, or directly from the website). (B) Localization of A645 from the 25S rRNA in *S. cerevisiae* (PDB 6QTZ), that is one key single RNA modification site listed in Bailey *et al.* (16).



**Figure 5.** Examples of applications of the *3D superimposition* tool. (A) Superimposition of a 90-residue range from proteins uL4 in *H. sapiens* (PDB 4UG0) and *E. coli* (PDB 5AF1), showing the existence of an extended loop at the eukaryotic exit tunnel (14) (see also Supplementary file 3). (B) Superimposition of protein uL23 (same structures as in A), showing an extra segment in bacteria, replaced by protein eL39 in eukaryote (14). The eukaryotic proteins are in orange whereas the bacterial are in white and the residues forming the extended loops are highlighted.

a list of residues present in the vicinity of the ligand in the original structure (with a cutoff distance set at 5 Å). It is then transposed to the strands of the target structure via Biopython's `Bio.pairwise2` module (13).

## API

RiboXYZ also allows programmatic access to the data with a dedicated application programming interface (API). The API is partitioned into two endpoints, one for statically-served, downloadable files and the other for querying semantic data from the database. There are multiple endpoints available, with a more detailed description provided in the user manual.

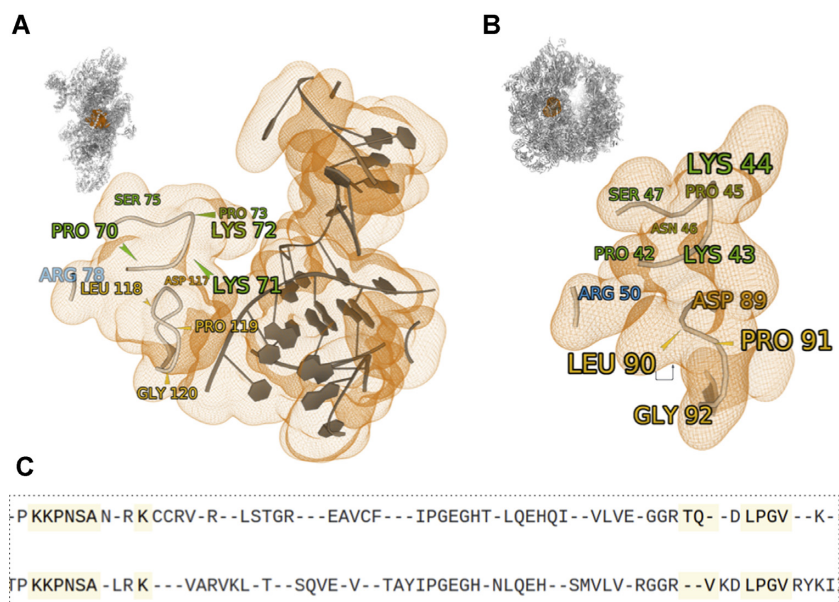
## EXAMPLES AND APPLICATIONS

We now provide various examples of applications from RiboXYZ. Some of these examples refer to recent publications and results, which the user can reproduce and interpret by using the database. A more detailed description of

the steps associated with each example is provided as tutorials in the database user manual, with videos as Supplementary files showing screen recordings. Both the user manual and videos are also directly accessible from the website.

## Residue selection and highlight

To visualize a specific structure or any region of a protein or RNA site of interest, the user can find and select an adequate structure from the database and use the visualization tool, as illustrated in the next two examples. First, we visualize in Figure 4A the Peptidyl-Transfer-Center (PTC) region of the ribosome in human, by highlighting the sites 2445–2452 from the *E. coli* 23S rRNA (14), with all steps to obtain the figure accessible detailed in our online user manual, or by video in Supplementary file 2, also available from the website. Note that as the PTC is a well-conserved rRNA region (15), the identification of the rRNA site can be done for any species upon aligning its 23S/25S/28S rRNA sequence with that of another well known organism, e.g. human in eukaryote or *E. coli* in bacteria (as further detailed in an-



**Figure 6.** Example of application of the *Ligands/Binding Sites* tool. We use the tool to highlight in (A) the structure of PDB 6RW4 (from *T. thermophilus*) with antibiotics streptomycin highlighted, and in (B) the prediction of the antibiotic's binding site in *Mycobacterium tuberculosis*, using the PDB 5ZEP structure (see also (21)). The residues successfully mapped from source structure 6RW4 to target structure 5ZEP are highlighted. (C) The sequence alignment between the two *uS12* proteins in source and target, on which the prediction is based, as provided by RiboXYZ (using the 'Inspect Prediction' button). See also Supplementary file 4.

other example). In our second example (Figure 4(B)), we visualize one of the single rRNA modifications sites (7 sites in 18S rRNA and 73 sites in 25S rRNA) studied in Bailey *et al.*, that were recently found susceptible to affect translation, ribosome biogenesis, and pre-mRNA splicing (16). The figure illustrates the specific location of site 645 of the 25S rRNA, mentioned in Figure 1 of Bailey *et al.*

### 3D superimposition and comparison of homologous proteins

The 3D superimposition tool of RiboXYZ can be used to compare and visualize structural differences of the ribosomal components across different species or conformational variations of the ribosome (as described in the *Visualization Tools* section). We applied this tool in Figure 5, where we focused on two ribosomal proteins: uL4 and uL23. Both proteins are also part of the ribosome exit tunnel, a sub-compartment of the ribosome that contains the nascent polypeptide chain (17). In particular, comparing and visualizing these two proteins allow to explain and interpret the main differences found in the geometry of the tunnel between eukaryotes and prokaryotes (14), that also impact the nascent chain folding and its escape from the tunnel (18,19). In Figure 5(A), we illustrate the superimposition obtained for protein uL4 with ribosomes from human (PDB 4UG0) and *E. coli* (PDB 5AFI) (with all steps to obtain the figure detailed in our online user manual, or by video in Supplementary file 3, also available from the website). In particular, the alignment reveals an extension of the loop region of the exit tunnel in human, as mentioned in Figure 4 of Dao Duc *et al.* (14). Similar superimpositions can be performed to compare structures across domains of life and species, by finding representative structure in the database (e.g. PDB 4V9F in archaea). In Figure 5B, we show another compar-

ison between the eukaryotic and prokaryotic ribosomes in uL23, which again shows the existence of another loop in bacteria, that is absent in eukaryotes (as it is replaced by eL39m see also Figure 5 from (14)). We also note that in these two examples, we improved the accuracy of the superimposition by adjusting the regions to align (which can be done using a dedicated bar tool or by manually entering the positions).

### Visualization and prediction of antibiotic binding site

The *Ligands/Binding Sites* tool allows to compare and use structures that include ligands (e.g. m/tRNA, translation factors, antibiotics etc.). The user can for example use it to predict and visualize the binding sites of the same ligand across species (assuming that the binding site remains the same), or compare structures from the same organism with and without any given ligand. The prediction of the ligand binding site for an other structure can be useful for molecular docking studies, and is notably relevant for the study of macrolide antibiotics, which constitute an important class of ligands that were found in various structural studies to target different locations of the ribosome, such as the PTC or the exit tunnel (20). Yet, the structure of a ribosome bound with a specific antibiotic is in general not available for a species that does not belong to the most studied structures (see also Table 1). For example, the effect of the antibiotic streptomycin on *Mycobacterium tuberculosis* was recently investigated, but with a structure from *Thermus thermophilus* used to visualize the ribosome ligand complex (21), as it is the only available structure of the ribosome bound with streptomycin. We hence first show, in Figure 6A, how one can use the *Ligand/Binding sites* tool to first visualize the binding site of streptomycin, in the riboso-

mal structure of the human mitochondrion (PDB 6RW5), and subsequently visualize it in the ribosome of *M. smegmatis* using PDB 5ZEP (Figure 6B). To do so, our tool first locates the binding site in 6RW5 at the vicinity of several residues of uS12 (see *Visualization Tools* section), and then produces a list of homologous sites in *M. tuberculosis* from doing pairwise sequence alignment (as described in the *Visualization tools* section), with the result of the alignment shown when clicking on the 'Inspect Prediction' button. All the steps to obtain the figure are detailed in our User manual online and accessible by video in Supplementary file 4, or directly from the website.

### Using curated structures and API

We finally mention two examples of use of the database and its edited structures. First, the database can facilitate the localization of the PTC and exit tunnel for a given set of species. Upon downloading any structure of interest from the database, the user can take advantage of the common ontology used for the RNA and ribosomal proteins, and programmatically select the nucleotides located at the PTC. The tutorial section of our user manual contains a step-by-step description of how to run this procedure, with its implementation as a Python3 script available as Supplementary File 5. In particular, this procedure can potentially automate the manual processing done in our previous paper (14), to initialize an algorithm extracting the exit tunnel location (22). While this comparison was done for twenty structures in (14), using RiboXYZ would allow in principle to extend the localization of the tunnel to all the structures that contain the exit tunnel.

Secondly, the webserver ProteoVision (23) provides an example of how RiboXYZ's API can be used for any application that requires a large number of curated structures. The DESIRE mode of ProteoVision dedicated to evolutionary analysis and visualization of ribosomal proteins (rProteins) across the Tree of Life, provides an integration of the multiple sequence alignments of rProteins with their topology diagrams, and the selected ribosomal 3D structure. The integration is facilitated by use of API annotations of rProteins from RiboXYZ that resolves the relationship between the protein name according to the current notation (5) and provides a list of all available PDB structures containing the selected protein name. The details of integration are available in the ProteoVision's github repository.

## DISCUSSION

Here we present RiboXYZ, a database that provides a simplified and curated access to ribosome structures with several visualisation tools. The database provides a unified nomenclature of all ribosomal components across all structures available in PDB. While the RCSB PDB server has recently integrated more tools for visualization and comparison (24), the interface of RiboXYZ was designed to be light and intuitive, in comparison with the more dense PDB interface. As a result, we hope that non-specialists in structural biology can easily get access to structures, and visualize any result that can relate with some aspect of the ribosome structure, as illustrated in our examples and appli-

cations (14,16,21). In this regard, a fundamental contribution of our database is that our data structure disambiguates the identification of the ribosomal components, that was so far hindered by the use of arbitrary ID's in deposited structures and other naming convention outside from the field of structural biology (6,7). Previous efforts to integrate ribosome structural data were made to build databases and interfaces for 3D alignment (25), or jointly visualize 1D, 2D and 3D structures of the ribosome (26). While these databases are not available anymore and/or do not scale up to the current amount of available structures (4), the past year has interestingly seen the release of new applications with dedicated online server and database, that perform some advanced analysis of ribosome structures and proteins using a large dataset of structures available (23,27). For instance, Radtool (27) evaluates the relative rotation of the two ribosomal subunits across all structures available from the PDB. ProteoVision, which was earlier mentioned as it uses RiboXYZ's API allows to perform an evolutionary analysis and visualization of ribosomal proteins over a large number of species in the tree of life. Interestingly, ProteoVision jointly integrates a 3D viewer (9) using structures from the PDB like RiboXYZ, which demonstrates how structure visualization can help interpret results derived from sequence analysis.

The database is also built to be augmented and maintained over time, with the release of new structures, and the potential integration or crosslinking with other databases to improve its coverage and annotations (12). As RiboXYZ applies a generic framework to process and describe all the structures, some more effort can be done in the future to provide more specific annotations and tools to cover some fundamental aspects of the ribosome that were recently addressed in structural studies, including the conformational heterogeneity (4), assembly pathways (28), and other functions of the ribosome (e.g. antibiotic resistance (20), nascent chain interactions (17)). In this regard, a first step to enhancing the capabilities of the database would be to provide a comprehensive and direct access to all non-ribosomal proteins from the current structures, that are so far only indirectly accessible using our Ligand/Binding sites tool. Additionally, we plan to include all the structures containing the ribosomal components solved in isolation (as single chains). While the tools produce satisfactory results in our examples, we also plan for further improvement and alternative options. For instance, using pairwise structural alignment methods (29), or leveraging curated seed-sequences for multiple sequence alignment (23) can be done as an alternative to the sequence alignment, currently used in the superimposition and binding site tools. Overall, these potential modifications can be easily implemented, given the modular architecture of the database. More generally, while the current tools mostly apply to single structure for visualization, or pairs of structures for superimposition and binding sites prediction, it would be interesting to fully leverage the diversity of structures in our database by designing tools that perform multiple and high-throughput analysis across the database.

Finally, we note that categorizing the PDB is becoming increasingly necessary, to maintain uniformity of access against the proliferation of heterogeneous data. RiboXYZ

addresses this issue, for a subset that we augmented with specific domain knowledge and ‘data-oriented’ programming. As such, it is a proof of concept of a model that would be interesting to apply to other important classes of molecules.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Loren Williams’s group for discussion.

## FUNDING

NSERC Discovery Grant [PG 22R3468]; UBC STAIR grant; Peter Wall Institute for Advanced Studies and Biotalent Canada (to K.D.D. and A.K.); National Aeronautics and Space Administration [80NSSC18K1139 to A.S.P.]. Funding for open access charge: NSERC Discovery Grant [PG 22R3468].

*Conflict of interest statement.* None declared.

## REFERENCES

- Ramakrishnan, V. (2018) Gene Machine: the race to decipher the secrets of the ribosome, *Basic Books*.
- Earl, L.A., Falconieri, V., Milne, J.L. and Subramaniam, S. (2017) Cryo-EM: beyond the microscope. *Curr. Opin. Struct. Biol.*, **46**, 71–78.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
- Poitevin, F., Kushner, A., Li, X. and Dao Duc, K. (2020) Structural heterogeneities of the ribosome: new frontiers and opportunities for Cryo-EM. *Molecules*, **25**, 4262.
- Ban, N., Beckmann, R., Cate, J.H., Dinman, J.D., Dragon, F., Ellis, S.R., Lafontaine, D.L., Lindahl, L., Liljas, A., Lipton, J.M. *et al.* (2014) A new system for naming ribosomal proteins. *Curr. Opin. Struct. Biol.*, **24**, 165–169.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C., Paladin, L., Raj, S., Richardson, L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
- The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
- Van De Waterbeemd, M., Tamara, S., Fort, K.L., Damoc, E., Franc, V., Bieri, P., Itten, M., Makarov, A., Ban, N. and Heck, A.J. (2018) Dissecting ribosomal particles throughout the kingdoms of life using advanced hybrid mass spectrometry methods. *Nat. Commun.*, **9**, 1–12.
- Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J. and Rose, A.S. (2021) Mol\* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
- Rose, Y., Duarte, J.M., Lowe, R., Segura, J., Bi, C., Bhikadiya, C., Chen, L., Rose, A.S., Bittrich, S., Burley, S.K. *et al.* (2021) RCSB Protein Data Bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *J. Mol. Biol.*, **433**, 166704.
- The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
- Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
- Chapman, B. and Chang, J. (2000) Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, **20**, 15–19.
- Dao Duc, K., Batra, S.S., Bhattacharya, N., Cate, J.H. and Song, Y.S. (2019) Differences in the path to exit the ribosome across the three domains of life. *Nucleic Acids Res.*, **47**, 4198–4210.
- Doris, S.M., Smith, D.R., Beamesderfer, J.N., Raphael, B.J., Nathanson, J.A. and Gerbi, S.A. (2015) Universal and domain-specific sequences in 23S–28S ribosomal RNA identified by computational phylogenetics. *RNA*, **21**, 1719–1730.
- Bailey IV, A.D., Talkish, J., Ding, H., Igel, H.A., Duran, A., Mantripragada, S., Paten, B. and Ares, M. Jr. (2022) Concerted modification of nucleotides at functional centers of the ribosome revealed by single-molecule RNA modification profiling. *Elife*, **11**, e76562.
- Ito, K. (2014) *Regulatory Nascent Polypeptides*. Springer.
- Kudva, R., Tian, P., Pardo-Avila, F., Carroni, M., Best, R.B., Bernstein, H.D. and Von Heijne, G. (2018) The shape of the bacterial ribosome exit tunnel affects cotranslational protein folding. *Elife*, **7**, e36326.
- Chwastyk, M. and Cieplak, M. (2021) Nascent folding of proteins across the three domains of life. *Front. Mol. Biosci.*, **8**, 508.
- Vázquez-Laslop, N. and Mankin, A.S. (2018) How macrolide antibiotics work. *Trends Biochem. Sci.*, **43**, 668–684.
- Li, S., Poulton, N.C., Chang, J.S., Azadian, Z.A., DeJesus, M.A., Ruecker, N., Zimmerman, M.D., Eckardt, K.A., Bosch, B., Engelhart, C.A. *et al.* (2022) CRISPRi chemical genetics and comparative genomics identify genes mediating drug potency in *Mycobacterium tuberculosis*. *Nat. Microbiol.*, **7**, 776–779.
- Sehnal, D., Vařeková, R.S., 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. Cheminform.*, **5**, 39.
- Penev, P.I., McCann, H.M., Meade, C.D., Alvarez-Carreño, C., Maddala, A., Bernier, C.R., Chivukula, V.L., Ahmad, M., Gulen, B., Sharma, A. *et al.* (2021) ProteoVision: web server for advanced visualization of ribosomal proteins. *Nucleic Acids Res.*, **49**, W578–W588.
- Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichtlow, G.V., Duarte, J.M., Dutta, S., Fayazi, M., Feng, Z. *et al.* (2022) RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. *Protein Sci.*, **31**, 187–208.
- Jarasch, A., Dziuk, P., Becker, T., Armache, J.-P., Hauser, A., Wilson, D.N. and Beckmann, R. (2012) The DARC site: a database of aligned ribosomal complexes. *Nucleic Acids Res.*, **40**, D495–D500.
- Bernier, C.R., Petrov, A.S., Waterbury, C.C., Jett, J., Li, F., Freil, L.E., Xiong, X., Wang, L., Migliozzi, B.L., Hershkovits, E. *et al.* (2014) RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.*, **169**, 195–207.
- Hassan, A., Byju, S., Freitas, F.C., Roc, C., Pender, N., Nguyen, K., Kimbrough, E.M., Mattingly, J., Gonzalez, R.L., de Oliveira, R.J. *et al.* (2022) Ratchet, swivel, tilt and roll: a complete description of subunit rotation in the ribosome. bioRxiv doi: <https://doi.org/10.1101/2022.06.22.497108>, 22 June 2022, preprint: not peer reviewed.
- Klinge, S. and Woolford, J.L. (2019) Ribosome assembly coming into focus. *Nat. Rev. Mol. Cell Biol.*, **20**, 116–131.
- Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. and Godzik, A. (2020) FATCAT 2.0: Towards a better understanding of the structural diversity of proteins. *Nucleic Acids Res.*, **48**, W60–W64.