Analysis and illustration of primary and secondary structures of ribosomal RNA and ribosomal proteins

Faculty Member #1 (Adviser):

Printed:_____

Signature:_____

Faculty Member #2:

Printed:_____

Signature:_____

INTRODUCTION

RiboVision is a collection of applications housed on servers at the Georgia Institute of Technology which serves to facilitate the development of publication-quality diagrams of ribosomal RNA (rRNA) and ribosomal protein (rProtein) structures (Petrov et. al, 2014). In particular, RiboVision seeks to promote analysis of key properties of rRNA and rProteins in primary, secondary, and tertiary structures.

As key semantides (ubiquitous macromolecules which carry genetic equivalent to the information intrinsic to DNA molecules and may be used by comparison to inform phylogenetic relationships), comparison of the primary and secondary structures of 16S and 18S RNA allows for the phylogenetic comparison of prokaryotic species and eukaryotic species, respectively (Fuerst, 2001). Sequence alignments are housed on the RiboVision server and stored in a MySQL database. Over the next two semesters, major improvements will be made to the server resulting in the newest edition, RiboVision3, which will feature improvements over the preceding RiboVision2 including the integration of XRNA, a program responsible for the generation of rRNA secondary structures and their exportation of their data into common computer-file formats (CSV, SVG, PDF, etc.) and the PDB Topology Viewer, a program responsible for production of protein secondary structures and their exportation into SVG image files.

The core functionality of XRNA - demonstration and editing tools of rRNA secondary structures needs to be iterated upon to allow for a more diverse set of purposes, including processing of high-quality hand-edited images into formats which are compatible with on-server management and conversion into formats native to web browsers.

In addition, major improvements need to be made to the architecture of queries made to the server's database backend; these include support for filtration of rRNA and rProtein primary structures based on taxonomy - by superkingdom, phyla, and more. These filters improve the application's usability and navigability in the creation of sequence alignments. Once made, these improvements shall facilitate the development of sequence alignments leading to generation of phylogenetic trees of multiple, potentially distantly related species.

These improvements constitute a major overhaul of RiboVision; collectively, these planned services as supported by in-house servers allow for researchers both internal and external to the Williams lab to generate publication-quality diagrams and analyses of rRNA and rProtein primary, secondary, and tertiary structures or to compare the relative relation of species by use of the ribosome as a semantide.

LITERATURE REVIEW

As a vital component of every cell in the process of transcription, ribosomes serve as a key semantide for research investigating the evolution of life on earth. A semantide serves as an information-carrying universal cellular molecular structure which mutates at a reasonably slow rate, thereby allowing for analytical determination of the evolutionary relationships of biological organisms which house these molecular machines; 16S rRNA is one of this class of macromolecules [1]. Traditionally, the primary structure - the sequence of nucleotide sub-molecules (monomers) of ribosomes - was used as the primary determining factor in the quantitation of variation between ribosomal structures [1]. As precision in determination of ribosome structures' tertiary structures has improved, the ability of researchers to utilize ribosomal secondary structure in coordination with ribosomal tertiary-structure differences in their analysis has led to more precise evaluation of phylogenetic relationships [2].

Many prominent studies contend that the central regions of the ribosome are the oldest regions which are most fundamental to the function of ancient ribosomes [3]. These assertions are partially made on the confidence of the analysis of aforementioned precisely determined ribosomal tertiary structures in concordance with associated secondary structures. The degree of conservation of ribosome regions between distantly related species are inversely correlated to the distance (atomic root-mean-squared distance) from the ribosome's peptidyl transfer center; these observations were made during analysis of 23S rRNA [3]. As a result, analysis the expanding regions of the ribosome as a function of species variation may be used to elucidate phylogenetic relationships of even distantly related species. An important caveat is that ribosomal secondary

structure conservation appears to be more significant to the conservation of ribosome function than conservation of ribosomal nucleic-acid sequences; ribosomal primary structures therefore must be analyzed in parallel with their secondary structure in order to properly predict phylogeny [4].

Some of the work being done in the Williams lab is the production of a suite of tools for the graphical depiction and analysis of differences between ribosomal structures (primary, secondary, and tertiary); this suite of tools (known as RiboVision) is available on an in-house server and provides public access to researchers' usage of these tools in their publications [5]. Generally, RiboVision serves to provide its users with high-quality images of ribosomal structures, using common color pallets and high-resolution (publication-quality) image production in PNG and SVG formats. These images are produced through structure data uploaded to the RiboVision server and are used both in the Williams lab and in the general discourse of ribosomal structure to convey arguments concerning discrepancies of known vs. proposed ribosomal structures and between known (or commonly accepted) ribosomal structures of potentially distantly related biological species. RiboVision also serves to facilitate the production of sequence alignments, which are the comparison of rRNA primary sequences. As described earlier, the primary structures do not adequately inform phylogenetic determinations as deletions or insertions of even large portions of sequences may occur over the course of ribosomal evolution while maintaining consistent (or at least analogous) ribosomal secondary and tertiary structures [6]. Quite deliberately, RiboVision affords its users conjunctive depictions

of ribosomal RNA and ribosomal proteins alike, allowing for rapid and high-quality analysis of the complete structure of the ribosome.

METHODOLOGY:

Frameworks Used in the RiboVision server:

The major frameworks incorporated into the novel functionality of RiboVision3 include Django, XRNA, and Plotly.

The components of the in-progress version of RiboVision are being assembled independently. For the modeling and analysis of multiple-sequence alignments, a custom web server is being implemented; this site is backed by a database housed at the Georgia Tech Williams Lab. MySQL scripts facilitate the transfer of gene sequences and phylogenetic data between the server and client machines. The URL structures and interconnections of the site are managed by the Django Framework. Django provides a baseline level of integration between URLs. Building on this, Javascript scripts ferry commands and data between the new RiboVision web pages, while Python scripts direct data to other programs (like MAFFT) to generate user-directed multiple sequence alignments in concert with gene sequences scraped from the European Bioinformatics Institute (EBI). These processes are facilitated by the server's Linux operating system.

Python is also used to calculate important metrics of gene sequence similarity, including per-residue-index twin-cons/shannon entropy. Note the frequency of specific amino acids or nucleotides is sometimes used as the probability metric incorporated into the entropy calculations. Alternatively, amino acids are sometimes bracketed into classifications and the frequency of residues within those classes are instead used to make entropic analyses; example classifications include hydrophobicity, polarity, molecular weights, etc. Custom algorithms are currently being developed at the Williams lab to provide custom classifications, and custom phylogenetic similarity analyses arise as a direct result.

Another important tool being developed for the RiboVision server is the display of custom data uploaded by the site's users. These include RNA secondary structures in CSV file formats and multiple-sequence alignments (.fa FASTA files) made independently of RiboVision. These custom data sets are then displayed on custom HTML pages as backed by the PLOTLY program; PLOTLY is directly compatible with Javascript, allowing for rapid integration of data

and scripts which sometimes originate within the server and are sometimes uploaded directly by the site's users.

RNA Structure Analyses:

One major addition to the newest version of RiboVision is the integration of the XRNA program. XRNA is a Java program originally compiled with Java 6. Along with newer versions of Java have come some important feature updates and improvements, including support for high-resolution screens in native Java packages. A number of new features have been added to XRNA; these include exporting the program's data to SVG image files and CSV files. These file conversions occur without loss of information.

In addition, some new algorithms have been added which allow for direct import of SVG files to XRNA's internal data. To accomplish this, a battery of geometric algorithms operate to cooperatively establish the relationships between primitive graphical elements in the SVG files. This is necessary, as many of the SVG files targeted for import into XRNA are not programmatically constructed. Instead, they are frequently touched up by human hand to improve the arrangement of residue labels by the XRNA program. It is therefore desirable to establish a script which tolerates error in placement of lines pointing in the general direction of nucleotide residues. In addition, the data of nucleotide text, label text, and label lines are segregated within different groups of the SVG file; these must therefore be carefully parsed to allow a Java algorithm to establish the relationships of SVG elements geometrically. These geometric algorithms include line-point distances, line segment-point distances, a custom algorithm for calculating the degree of overlap of directionality of line segments (incorporation of linear algebra) and some error correction algorithms.

A number of improvements have yet to be made to XRNA; a number of these include bug fixes, especially bug fixes that relate to inappropriate modification of XRNA internal data. The next most significant improvement to be made to XRNA is Java to Javascript transpilation, a process which allows for the creation of Javascript source code which seeks to establish equivalent Javascript code from Java source code. Once complete, this process will allow XRNA to run natively within RiboVision's servers. Additionally, this is important in that it can allow for

direct integration of automated processes, including the creation of publication-quality SVG images.

RESULTS:

The most significant task for the improvement of the XRNA program was the creation of an algorithm for seamless or close-to-seamless conversion of XRNA files to SVG files. By nature of SVG files as an image file, its data can be somewhat unstructured; regardless of the relatedness of adjacent elements as the eye would see (lines pointing from labels to nucleotides) SVG data needs not be grouped. Depending on the format applied and the algorithms used to produce the data, the lines, labels, and nucleotides could be completely scattered about the file without much rhyme or reason. Therefore, it was a major challenge to derive a method by which independent file elements could be computationally observed to be related to one another.

Multiple geometric metrics for spatial relatedness immediately come to mind when designing such an algorithm; minimizing distance, maximizing directedness of lines pointing between labels and nucleotides, etc. Ultimately, it was a combination of these metrics which allowed for computationally determined relatedness to approximate the way a human observer would naturally relate the images' disparate elements. Much trial and error took place to create a reliable algorithm, and even upon its completion, some error tolerance needed to be built in. This came in the form of calculating and subtracting the most common indexing error; as nucleotides labels are generally numbers, they could be parsed from text to integer form. Then, the difference between the label's integer value and the index of the first nucleotide was calculated. Some labels could not be appropriately placed despite the algorithm's best efforts. As a result, the random error was nullified, as the most common (modal) difference between a label's parsed value and the index of the first nucleotide was subtracted from the calculated per-label index. The algorithm's results are largely positive, with some rare errors.

The SVG files were otherwise improved by the addition of symbols to signify non-canonical base pairs. Hollow circles signify wobble base pairs while filled circles represent other non-canonical base pairs.

The most recent feature addition to the XRNA program is an improvement of the algorithm used to geometrically transform the bounds of XRNA nucleotide data to bounds specified by the RiboVision server; Now, XRNA will reliably produce CSV export files with correct orientation and bounding, even when the input XRNA file contains multiple molecules.

Other improvements made to the XRNA program include an improvement of the way XRNA represents its internal data; specifically, rather than representing label line segments as quadrilateral shapes straddling two points, two points now represent label lines. This was motivated by an observation of the previous line representation causing graphical artifacts.

RiboVision has also recently seen multiple feature additions, the most prominent of which is the selection of secondary-structure substructures; this allows for the conversion of per-sequence indices to per-alignment indices. Those alignment indices can be utilized to calculate per-alignment index entropy or propensity calculations, allowing for colored maps to be created; this feature is a significant tool added to our research partner's toolbox as they seek to create publication-quality diagrams of their data.

DISCUSSION:

The newest features added to the RiboVision server are integral to our research partners' projects, especially the production of publication-quality images (whether they are of secondary structures, alignment propensity or twincons entropy data, or other such per-residue data maps). The features directly aid researchers both within the Williams Lab and elsewhere as they strive to automate their research and perform high-throughput calculations.

New tools added to the Ribovision server such as the custom selection of secondary-structure substructures will allow our research partners to analyze their sequence alignments in greater detail, without resorting to utilizing a separate program; this enhances efficiency and can be engineered in such a way that automation of the overall process can be maintained in a single script. One example of this feature integration has already occurred within our lab: per-residue propensity calculations. Per-residue propensities can allow for investigation of evolutionary relationships between closely related genes (whether orthologs or paralogs) and can be a highly useful tool in the investigation of the function of proteins with regions of low conservation (i.e. highly variable regions of residues.)

Direct applications of the newest XRNA features allow for more efficient production of ribosomal secondary-structure maps, circumventing steps which were previously accomplished by manual student labor. These include the maintenance of XRNA files independently of the SVG images which they are logically equivalent to (they contain the same information). With the added functionality of import of SVG files, only one file needs to be maintained. Similarly, improved transformation of XRNA data into CSV format allows for researchers in our lab to house the same data on the RiboVision server in a more compact format.

CONCLUSION AND FUTURE WORK:

The tools added to the newest version of RiboVision focus on allowing for the creation of alignments with ease on the part of the user while maximizing the customizability afforded to them. These goals manifest themselves most prominently in the selection tools which are now available for selection of portions of secondary structures in coordination with alignment maps to generate highly customized colored secondary-structure maps. In addition, the tools for uploading user-provided fasta files serve as a direct alternative to the construction of database-source phylogenetic alignments as calculated on the RiboVision server.

The features added to the XRNA program practically make XRNA feature-complete; future steps for the RiboVision program are largely composed of maintenance of the GitHub repository (https://github.com/LDWLab/XRNA-GT), and bug fixes as they arise. Only one other major task is planned for ribovision: transpilation to Javascript code so that a Javascript version of XRNA can be directly embedded into the RiboVision server.

A recent application of the XRNA program and some of its newest features is published in "R2DT is a framework for predicting and visualizing RNA secondary structure using templates," Nature Communications, 2021, in press (B. Sweeny et al, 2020).

Further improvement to the SVG image conversion method can and should be made; very short nucleotide-bond lines can cause somewhat unpredictable results. For now, these lines are mostly left out of the reconstruction of XRNA files to prevent erroneous bonds being constructed; further analysis and algorithmic tweaking could lead to more robust handling of this edge case.

Future RiboVision projects will focus on providing new ways for the user to interact with data uploaded to the web server; new ways to select secondary-structure subsections need to be implemented, including per-domain, per-index range, etc. These selection methods need to be designed in such a way that the user experience is intuitive and uncomplicated, in keeping with the design philosophy of the pre-existing RiboVision features.

CITATIONS:

1. Wayne, L. G., R. C. Good, E. C. Bottger, R. Butler, M. Dorsch, T. Ezaki, W. Gross et al. "Semantide-and chemotaxonomy-based analyses of some problematic phenotypic clusters of slowly growing mycobacteria, a cooperative study of the International Working Group on Mycobacterial Taxonomy." *International Journal of Systematic and Evolutionary Microbiology* 46, no. 1 (1996): 280-297.
2. Jobe, Amy, Zheng Liu, Cristina Gutierrez-Vargas, and Joachim Frank. "New insights into ribosome structure and function." *Cold Spring Harbor perspectives in biology* 11, no. 1 (2019): a032615.
3. Hsiao, Chiaolong, Srividya Mohan, Benson K. Kalahar, and Loren Dean Williams. "Peeling the onion: ribosomes are ancient molecular fossils." *Molecular Biology and Evolution* 26, no. 11 (2009): 2415-2425.
4. Kjer, Karl M. "Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs." *Molecular Phylogenetics and Evolution* 4, no. 3 (1995): 314-330.
5. Bernier, Chad R., Anton S. Petrov, Chris C. Waterbury, James Jett, Fengbo Li, Larry E. Freil, Xiao Xiong et al. "RiboVision suite for visualization and analysis of ribosomes." *Faraday discussions* 169 (2014): 195-207.
6. Ben-Shem, Adam, Nicolas Garreau de Loubresse, Sergey Melnikov, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. "The structure of the eukaryotic ribosome at 3.0 Å resolution." *Science* 334, no. 6062 (2011): 1524-1529.
7. Fuerst, John A. "Semantides and modern bacterial systematics." *e LS* (2001).
8. Sweeney, Blake A., David Hoksza, Eric P. Nawrocki, Carlos Eduardo Ribas, Fábio Madeira, Jamie J. Cannone, Robin R. Gutell et al. "R2DT: computational framework for template-based RNA secondary structure visualisation across non-coding RNA types." *bioRxiv* (2020).