Visualizing ncRNA Structural Evolution with Arc Diagrams

Alyssa Tsiros, Lane Harrison
Worcester Polytechnic Institute

Abstract

The second challenge of this year's BioVis Symposium Design Contest was to create a visualization of noncoding RNA (ncRNA) structural evolution of the human accelerated region 1 (HAR1) gene in ancestral, denisovan, and human sequences. The commonly used chart types for this data, dot plots and node-link style graphs, do not support direct comparison of base pairings among the three structures. We propose a redesign that uses arc diagram visualization techniques to highlight conserved or otherwise evolved base pairings. These diagrams support the placement of unique and prevalent base pairings along a common scale and allow mapping task-specific features to color, providing viewers with a more direct means to identify differences in the structure predictions.

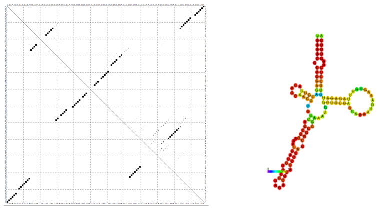
Keywords: Structure prediction, noncoding RNA, structural evolution, arc diagrams.

1. Introduction

There are several RNAs transcribed by the human genome known as non-coding RNAs, or ncRNAs, that do not code for proteins but still impact cellular processes. Until recent years, it was widely accepted that the most pertinent genetic information was obtained from proteins [4]. Evidence suggests, however, that the majority of mammal genomes is transcribed into ncRNAs [4]. Such evidence has created a need to classify ncRNAs, where each class has been shown to have a characteristic secondary structure [2]. Moreover, small changes in a primary RNA sequence can impact this structure and overall function, motivating molecular and structural biologists to compare secondary structures of evolved sequences.

To facilitate such comparisons, computational biologists have developed techniques to predict secondary structures from primary RNA sequences. These secondary structures are determined by a collection of predicted base pairs that minimize the free energy of the fold [6], which is also known as the minimum free energy (MFE) structure. The visualization of this secondary structure typically comprises a "dot-plot", or a grid of possible base pairs (or non-pairs) at each nucleotide residue position [5], along with a secondary structure node-link style graph of the MFE structure (see Figure 1).

While these dot-plots and secondary structure graphs are suitable visualizations for analyzing single genes, there remains a need for techniques that support the comparison of more than one set of genes along an evolution sequence. Such comparisons are difficult with the current dot plot and graph techniques for several reasons. One source of difficulty is that the dot plots are are arranged diagonally, making side-by-side comparison difficult. Another difficulty stems from the individual dots being small. Small dots emphasize large patterns over individual links, reminiscent of results on user performance with network visualization using matrix diagrams, which perform worse than node-link diagrams on tasks involving individual links [3]. These issues with the dot plot representation make it difficult for the viewer to connect features in the dot plot with the secondary structure graph.



1. Dot-plot showing probabilities of base pair bindings, and secondary structure graph [1].

Our goal in the redesign was to identify a representation that both emphasizes individual base pairings and supports the mental connection between base pairings and the secondary structure graphs.

2. Methods

The structure of the pairing data includes multiple sets of two-dimensional probability matrices, yielding a large design space for new visual representations. To narrow the design space and better align the possible visual representations with specific tasks, we generated several design alternatives and conducted semi-structured interviews with biologists and visualization practitioners.

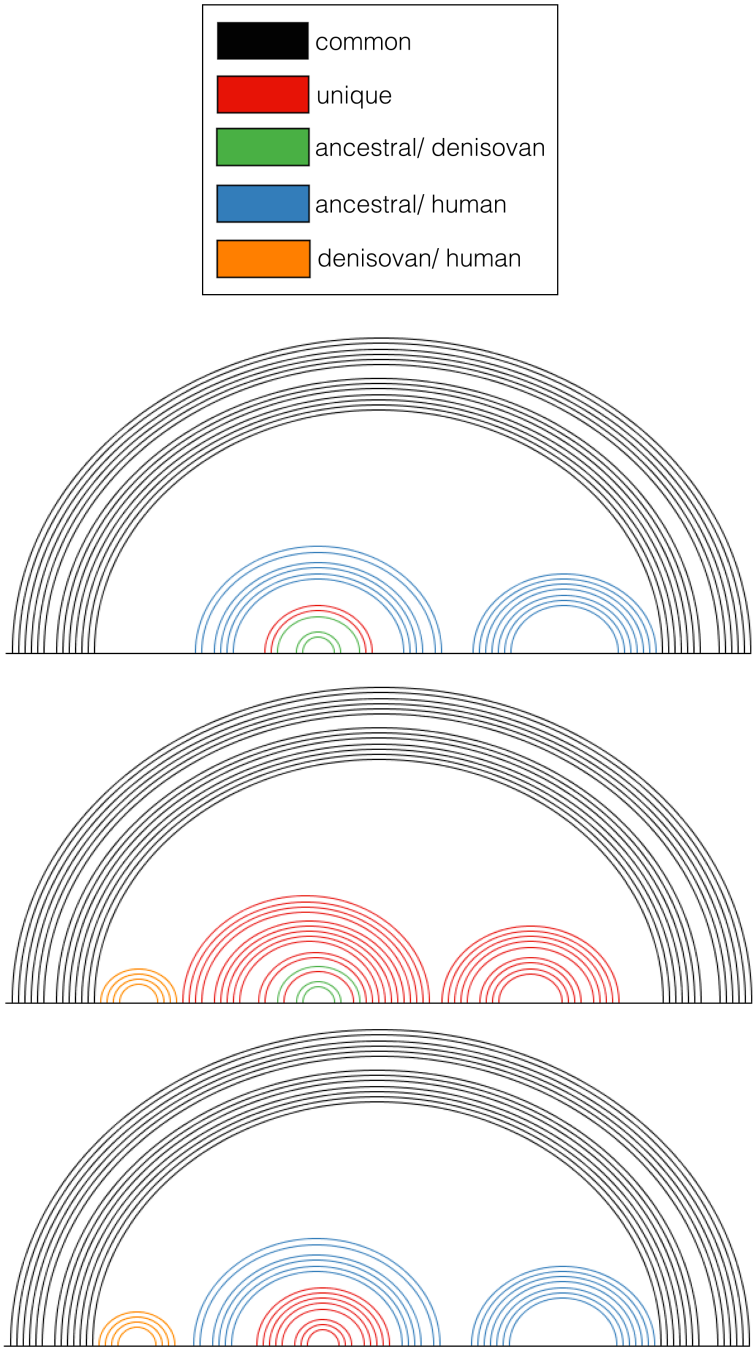
Interviewees included bioinformatics and computer science graduate students, as well as professors in computer science, evolutionary biology, and genetics. After being presented with an overview of the dataset and challenge, all were asked to comment on the BioVis Symposium design. The general consensus was that the new design should still clearly identify the different stems of the secondary structure prediction, while still supporting direct comparison of how the base pairings evolve from ancestral to human. Interviewees were then shown several possible redesigns, including circle plots, mountain plots, tree plots, and arc diagrams. Throughout the interviews, the arc diagrams were consistently the most popular choice, and therefore we pursued them further in our redesign.

Our redesign inspired by arc diagrams is titled "Relative distance versus minimum free energy base pairing for ancestral, denisovan, and human accelerated region 1 genes". Each of the provided sequences have unique and common MFE structure base pairings. To highlight these pairings, color is assigned five features including unique pairings, common pairings among ancestral and denisovan, common pairings among ancestral and human, common pairings among denisovan and human, and common among all three.

atsiros@wpi.edu
ltharrison@wpi.edu

The first, middle, and last positions of the sequence are indicated in the figure. Moreover, relative distance is calculated where k_1 and k_2 are the first and second positions of a MFE binding, respectively and L is the length of the sequence:

$$(| k_1 - k_2 | + 1) / L$$



2. Arc diagram redesign submission to the BioVis Symposium design challenge #2.

The graphs display half of an ellipse between two binding positions. Nothing is displayed if a position does not bind. The black pairings are conserved among ancestral, denisovan, and human genes. The green pairings are conserved from ancestral to denisovan. The blue pairings are common among ancestral and human, the orange pairings are conserved from denisovan to human, and finally, unique pairings are shown in red.

Compared to the original dot plots, arc diagrams provide more emphasis to visual features that are needed for the task of analyzing evolution sequences. Individual pairings are now represented as arcs instead of individual dots, creating a larger visual feature and one that provides more room for the use of color. Arcs also emphasize clusters of pairings, which help the viewer make connections to prevalent features (e.g. loops and chains) in secondary structure graph.

At the same time, there are limitations to the arc diagram approach. One is that all arcs are supposed to be "equal" in importance, but the nature of arc diagrams creates larger arcs for the pairings at the near and far of the graph, and smaller arcs for those towards the center. Future work in this area should evaluate such limitations and strengths of arc diagrams against other representations in a task-based setting. Other work could further explore the design space offered by arc diagrams, particularly for representing uncertainty.

References

1. (2015). Design challenge. 5th Symposium on Biological Data Visualization. Retrieved from http://www.biovis.net/year/2015/design-contest
2. Arora, A., Panwar, B., Raghava, G. PS. (2014). Prediction and classification of ncRNAs using structural information. BMC Genomics, 2014 (15): 127. Retrieved from http://www.biomedcentral.com/1471-2164/15/127
3. Ghoniem, M., Fekete, J., and Castagliola, P. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. IEEE (2004), 17–24.
4. Makunin, I. V., Mattick, J. S. (2006) Non-coding RNA. Human Molecular Genetics, 15 (suppl 1): R17 - R29. Retrieved from http://hmg.oxfordjournals.org/content/15/suppl_1/R17.short
5. The W.C. Ray Lab. (2015, January 22). Understanding RNA folding energy dot-plots. [Video file]. Retrieved from https://www.youtube.com/watch?v=v1UbIUZ8k9o
6. Zuker, M. (1995). Prediction of RNA secondary structure by energy minimization. Washington University Institute for Biomedical Computing. Retrieved from http://mfold.rna.albany.edu/doc/old-mfold-manual/index.php