ARNA: Interactive Comparison and Alignment of RNA Secondary Structure

Gerald Gainant [*]
LaBRI UMR 5800
University of Bordeaux 1
FRANCE

David Auber [†]
LaBRI UMR 5800
University of Bordeaux 1
FRANCE

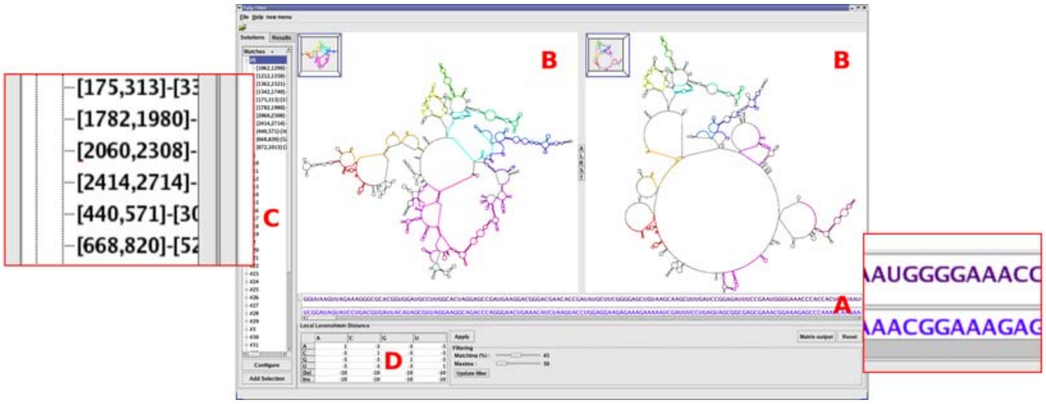

Figure 1: Comparing the secondary structure of two RNA sequences with ARNA.

Abstract

ARNA is an interactive visualization system that supports comparison and alignment of RNA secondary structure. We present a new approach to RNA alignment that exploits the complex structure of the Smith-Waterman local distance matrix, allowing people to explore the space of possible partial alignments to discover a good global solution. The modular software architecture separates the user interface from computation, allowing the possibility of incorporating different alignment algorithms into the same framework.

CR Categories: D.2.2 [Software Engineering]: Tools and Techniques—User interfaces; G.2.1 [Discrete Mathematics]: Combinatorics—Combinatorial algorithm; J.3 [Life and Medical Sciences]: Biology and genetics—;

Keywords: visualization, combinatorics, bioinformatics, graph drawing, sequence alignment, RNA

1 Introduction

ARNA is a new open-source system that provides support for biologists and bioinformaticians who need to compare the RNA secondary structures in two different organisms [3, 5]. The structure of RNA is often studied at three levels: the primary structure comprises the linear string of amino acids; the secondary structure is created when some amino acids in the sequence bond to others, forming two-dimensional structure; finally, the tertiary structure is formed when the sequence folds into a shape in three-dimensional space. ARNA always shows the primary RNA structure, and has side by side windows for showing the seconary or tertiary structures. ARNA is built within the open source framework Tulip [2][1].

Previous systems on secondary structure visualization, such as RnaViz [8] and Vienna [6], suffer from a lack of stability: small changes in the RNA primary structure may drastically change their drawing of its secondary structure. The scalable and stable tree-based drawing algorithm used in ARNA that uses a heuristic to locate and anchor quasi-isomorphic subgraphs shared between the two sequences is discussed in previous work [3]. We focus here on the problem of aligning RNA.

2 Aligning RNA

The problem of multiple sequence alignment has been well studied [4]. The RNA primary structure is a sequence R of nucleotides, represented as a word of length n on the alphabet $\{A,C,G,U\}$: $R = r_1 r_2 \ldots r_n$. Let W_R be the set of all the subwords of R. Let I_n be the set of all the sub-intervals on $[1,n]$:

$$I_n = \bigcup_{\substack{(i,j)\in[1,n]^2 \\ i\leq j}} [i,j]$$

Then $I_n \times I_m$ is the set of all possible matches between the two sequences of length n and m respectively. We would like find the similarity set, namely the subset that contains good matches.

Matrix Interpretation A prior approach by Smith and Waterman to finding similarities between two RNA sequences used a well-known local distance matrix built using the Levenshein distance metric [9]. They used this matrix to compute the longest possible match between the two sequences with the following heuristic: compute a *score* function $\varsigma : W_{R_0} \times W_{R_1} \longrightarrow \mathbb{R}$, find the maximum score, and then backtrack along the path that reached this score by reversing the computation. However, we noticed after implementing this algorithm in ARNA that this single longest possible match is far from the best match. We can instead find several good

[*] e-mail: gainant@labri.fr
[†] e-mail: auber@labri.fr

[1] `http://www.tulip-software.org`

shorter matches that when combined piecewise produce a far more precise fit for the data. We can do this using the information in the local distance matrix that is ignored in the Smith and Waterman algorithm. We show this information visually in Figure 2. Ths maximum score lies in the lower right corner of the matrix, and backtracking along its path corresponds to following the streak of that corner point's color up and to the left, roughly along the diagonal of the matrix. Looking at the matrix shows that there is a great deal of internal structure, with many paths that correspond to similarity ranges. ARNA allows users to interactively explore this rich internal structure of paths through the distance matrix; that is, to navigate through the space of matching ranges between the two sequences.

Filtering Controls (Fig. 1.D) We provide two simple controls: a "Maxima" slider to increase the range of scores allowed for the final path point, and a "Percentage Matching" slider to decrease the similarity set size by requiring the chosen percentage of sites in the sequence to match. We use a collection of heuristics to find the set of interesting paths through the matrix, including thresholds, geometric properties of paths, classification and reduction of the path set.

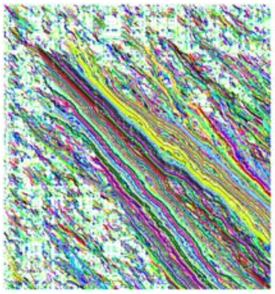

Figure 2: Path finding in the local distance matrix.

Filtering The set of paths that we extract from the matrix is still too large to be tractable for human exploration. We can make the task more manageable by forming groups of matches whose ranges do not overlap. Recall that $A_{n,m}$ is a subset of $I_n \times I_m$. We can recursively find a minimal set of classes C_k, a *convenient* subset of A, defined as :

$$C_k = \{(x_i, y_i)\} \qquad i \in [1,k],\ \ x_i \in I_n,\ \ y_i \in I_m$$

satisfying

$$x_i \cap x_j = y_i \cap y_j = \emptyset, \qquad \forall i \neq j$$

3 Interface

The graphical user interface of ARNA, shown in Figure 1, is based on the free Linux-based framework Qt[2].

Primary structures representations (Fig. 1.A) The primary structures of the two RNA sequences are displayed as text strings in the alphabet $\{A, C, G, U\}$ of bases. Users can pan using a scroll bar, and highlighting is linked to the selections in the other windows.

Secondary structures representations (Fig. 1.B) The two RNA secondary (or tertiary) structures are drawn side by side. The secondary structure drawing algorithm [3] respects the visual conventions used by biologists: structures such as stems, hairpins, bulges, interior loops, and multi-branch loops are shown in the way that matches manually created drawings of small sequences [7]. Moreover, the algorithm is stable across small changes to the primary structure and has successful heuristics for anchoring quasi-isomorphic substructures in the same place in each view.

ARNA includes the standard set of controls for navigation: zooming by dragging out a rubberband area or using the mouse wheel; panning by dragging, and recentering the view. The Tulip rendering algorithms provide realtime interaction even for very large datasets using progressive rendering [1]. The views can all be linked, so that actions performed on one view are mirrored in the other. For instance, navigation can be synchronized, and dragging a box over part of the secondary structure will also result in highlighting the primary structure view.

Alignment solutions and result area (Fig. 1.C) The leftmost panel contains the Solutions set of computed similarities, as filtered by the Maxima and Matching sliders, and the Results set of matches interactively chosen by the user. The user can investigate any of the Solutions set by clicking on the range in the Solutions window to see all the ranges in that group highlighted in different colors in the secondary window. The user can quickly decide which ranges to validate and which to discard. Over the course of exploration, the user adds the best solutions to the Results set of individual ranges. When the user has finished the exploration, the Results set constitutes the final alignment of the two RNA structures. When a range is placed in the Results set, the saturation of its coloring is decreased to indicate its resolved status.

4 Future work

We would like to improve the user interface, plugin framework, and matrix filtering. We would also like to support a wider variety of file formats. The system was architected with a modular separation between the interface and the computation, so that other approaches to alignment could be incorporated as plugins that use the same exploration interface. We would like to create several such alternative alignment plugins, based on feedback from our users.

5 Acknowledgements

We thank Tamara Munzner for her help with writing.

References

[1] D. Auber. Using Strahler numbers for real time visual exploration of huge graphs. In *International Conference on Computer Vision and Graphics*, volume 1-3 of *Journal of WSCG*, pages 56–69, 2002.

[2] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Software*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[3] D. Auber, M. Delest, J. P. Domenger, and S. Dulucq. Efficient drawing and comparaison of RNA secondary structure. *Tech report LaBRI RR-1325-04*, 2004.

[4] S.C. Chan, A.K.C. Wong., and D.K.Y. Chiu. A survey of multiple sequence comparaison methods. *Bull. Math. Biol.*, 4:563–598, 1992.

[5] S. Dulucq and L. Tichit. Rna secondary structures comparison: Edition and alignment algorithms. In *GASCom 2001 and Bijective Combinatorics*, pages 25–32, Sienne, 2001.

[6] I.L. Hofacker et al. Fast folding and comparison of RNA secondary structures. *Monatsheffe fr Chemie*, 125(167-188), 1994.

[7] P.B. Moore. Structural motifs in RNA. *Annu. Rev. Biochem.*, 68:287–300, 1999.

[8] P. De Rijk, J. Wuyts, and R. De Wachter. RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics*, 19:299–300, 2003.

[9] T.F. Smith and M.S. Waterman. Identification of common molecular sequences. *Mol. Biol.*, 147:195–197, 1981.

[2] http://www.trolltech.com