Bioinformatics, 36(11), 2020, 3578–3579
doi: 10.1093/bioinformatics/btaa147
Advance Access Publication Date: 3 March 2020
Applications Note

Structural bioinformatics

ASPRAlign: a tool for the alignment of RNA secondary structures with arbitrary pseudoknots

Michela Quadrini[1], **Luca Tesei**[2,*] and **Emanuela Merelli**[2]

[1]Department of Information Engineering, University of Padua, Padova 35131, Italy and [2]School of Sciences and Technology, University of Camerino, Camerino 62032, Italy

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on July 23, 2019; revised on February 23, 2020; editorial decision on February 24, 2020; accepted on February 26, 2020

Abstract

Summary: Current methods for comparing RNA secondary structures are based on tree representations and exploit edit distance or alignment algorithms. Most of them can only process structures without pseudoknots. To overcome this limitation, we introduce ASPRAlign, a Java tool that aligns particular algebraic tree representations of RNA. These trees neglect the primary sequence and can handle structures with arbitrary pseudoknots. A measure of comparison, called ASPRA distance, is computed with a worst-case time complexity of $\mathcal{O}(n^2)$ where n is the number of nucleotides of the longer structure.

Availability and implementation: ASPRAlign is implemented in Java and source code is released under the GNU GPLv3 license. Code and documentation are freely available at https://github.com/bdslab/aspralign.

Contact: luca.tesei@unicam.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA molecules have several functions that are mainly determined by their folded structure. Several methods for comparing RNA secondary structures have been introduced in the literature. Many of them are based on tree representations of the structures and exploit edit distance or alignment algorithms. An introduction to secondary structure comparison can be found in Schirmer *et al.* (2014) and a recent overview more focused on alignment is in Chiu and Chen (2017).

Pseudoknotted structures can be frequently found in RNA molecules and are associated with specific functions (Staple and Butcher, 2005). However, most of the existing approaches can process only pseudoknot-free structures or consider only particular subclasses of pseudoknots.

To overcome this limitation, we introduce ASPRAlign, a tool for comparing RNA secondary structures with arbitrary pseudoknots. ASPRAlign uses particular representations of secondary structures, called *Algebraic RNA Trees* and *Structural RNA Trees*, based on algebraic operators (Quadrini *et al.*, 2019). ASPRAlign computes a distance, called *ASPRA distance*, that only depends on the structure of the molecules and completely neglects the primary sequences. This is in line with the fact that several classes of RNA, such as messenger, transfer and ribosomal RNA, exhibit a highly conserved shape of the secondary structure, but little sequence similarity (Höchsmann *et al.*, 2004).

Given two secondary structures with arbitrary pseudoknots whose primary sequences are of length n_1 and n_2, ASPRAlign computes the ASPRA distance between them with a worst-case time complexity of $\mathcal{O}(n^2)$ where $n = \mathtt{max}(\mathrm{n}_1, \mathrm{n}_2)$. Its particular characteristics make ASPRAlign very efficient w.r.t. other existing methods. The comparison of our algorithm with those listed in Table 1 of Chiu and Chen (2017) shows that only ASPRAlign is quadratic while all the other alignment methods have a worst-case time complexity more than quadratic.

2 Implementation and usage

ASPRAlign is implemented in Java and runs on every Linux, Windows and Mac OS platform in which a Java SE Runtime Environment 8 (or higher) is installed. It includes the implementation of the Jiang *et al.* (1995) algorithm for tree alignment provided by the StatAlign software package (Arunapuram *et al.*, 2013; Novák *et al.*, 2008). Once installed, ASPRAlign can be called from the command line:

- `java -jar ASPRAlign.jar [options]` builds the Algebraic RNA Tree or the Structural RNA Tree of a given molecule or computes the ASPRA distance of two given molecules.
- `java -jar ASPRAlignWorkbench.jar [options]` compares all the molecules in a given input folder by computing the ASPRA distance between all possible pairs.

Option -h displays for both commands the full list of available options. Input RNA secondary structures with arbitrary

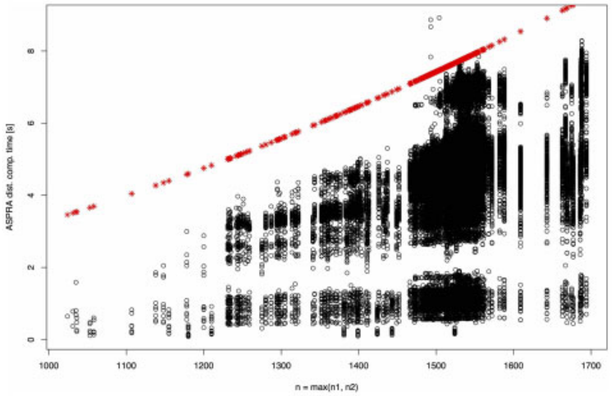

Fig. 1. Plot of execution times (black circles) in seconds for computing the ASPRA distance between pairs of pseudoknotted structures of length between 1000 and 1700. The red stars are the plot of $c \cdot n^2$ where c is a scaling constant and n is the length of the longer molecule in each execution. The four highest points above the red star line can be considered outliers due to operations of the operating system. (Color version of this figure is available at *Bioinformatics* online.)

pseudoknots can be given as extended dot-bracket notation or as arc-annotated sequence text files. The latter format is derived from the former by substituting the dot-bracket string with a list of pairs $(i_1, j_1); (i_2, j_2); \ldots; (i_m, j_m)$, where each index (i_k, j_k) belongs to the set $\{1, \ldots, n\}$, n is the length of the primary sequence, $i_k < j_k$ for all k and $m \leq \lfloor \frac{n}{2} \rfloor$. ASPRAlign needs a configuration file for specifying the cost of the elementary edit operations used by the alignment algorithm. The default file, called `ASPRAlign-config.txt`, is provided with the installation files. It can be edited or a different file can be specified. The default output format for Algebraic RNA Trees, Structural RNA Trees and Alignment Trees is a linearized tree of the form [‘`node-label`’ (`list-of-children`)]. For relatively small trees, LATEX output can be generated to produce a graphical representation of the tree in a pdf file. If `ASPRAlignWorkbench.jar` is used, the output is sent to CSV files containing the computed distances and additional information about the size of the molecules and the execution time. Examples of usage for both executable `jars` are reported in the Supplementary Material.

3 Features and performance

ASPRAlign is based on an algebraic representation of RNA secondary structures able to uniquely represent arbitrary pseudoknots using three operators: *concatenation*, *nesting* and *crossing*. The full definition of the operators can be found in Quadrini *et al.* (2019) and some explanatory examples are given in the Supplementary Material. The ASPRA distance is computed by aligning the Structural RNA Trees of the given molecules. These trees are an abstraction in which the primary sequence is ignored and only the structural part of the molecule is considered.

The parsing of input files containing molecules is performed by ASPRAlign in a time that is linear in the length n of the primary sequence. The construction of an Algebraic RNA Tree or of a Structural RNA Tree requires a number of scans of the primary sequence equal to the number of base pairs of the molecule. This requires a time complexity $\mathcal{O}\left(n + \sum_{k=1}^{m} (j_k - i_k)\right)$ where m is the number of base pairs and $(j_k - i_k)$ is the number of nucleotides enclosed by the kth base pair. In the worst-case we have that $m \simeq \frac{n}{2}$ and each $(j_k - i_k) \simeq n - 1$. Therefore, the time complexity can be equivalently assessed as $\mathcal{O}(n^2)$.

For aligning the Structural RNA Trees, ASPRAlign uses the implementation of the Jiang *et al.* (1995) algorithm whose time complexity is $\mathcal{O}(|t_1| \cdot |t_2| \cdot (\deg(t_1) + \deg(t_2))^2)$ where $|t|$ is the number of nodes of the tree t and $\deg(t)$ is the degree of the tree t, i.e. the maximum number of children of any node in the tree. Structural RNA Trees have degree 2 and the number of nodes is linear in the number of base pairs of the molecule. Thus, in the worst-case, the tree alignment and the computation of the ASPRA distance of two Structural RNA Trees with n_1 and n_2 nucleotides is performed in time $\mathcal{O}(\frac{n_1}{2} \cdot \frac{n_2}{2} \cdot 16) = \mathcal{O}(n_1 \cdot n_2)$. Overall, the cost of the construction of the trees and the computation of the distance is $\mathcal{O}(n_1^2 + n_2^2 + n_1 \cdot n_2) = \mathcal{O}(n^2)$ where $n = \max(n_1, n_2)$.

Figure 1 shows the plot of the execution times for comparing pairs of molecules that were processed using the ASPRAlign workbench facility. We considered a set of 460 molecules downloaded from RNAStrand (Andronescu *et al.*, 2008). Plotted times visually show the expected quadratic complexity w.r.t. the number of nucleotides of the longer molecule in the pair. More details are given in the Supplementary Material.

4 Conclusions

ASPRAlign is a Java tool for building Algebraic RNA Trees and Structural RNA Trees or to compute the ASPRA distance by aligning Structural RNA Trees of RNA secondary structures with arbitrary pseudoknots. With respect to other comparison methods, ASPRAlign performs in an efficient way ($\mathcal{O}(n^2)$). This is possible because of the algebraic representation of the molecules and the fact that the primary sequence is neglected. A graphical interface and a Web interface for ASPRAlign is under development. As future work, we want to extend the functionalities by adding a new measure based on edit distance and the possibility of aligning multiple molecules.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Andronescu,M. *et al.* (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Arunapuram,P. *et al.* (2013) StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. *Bioinformatics*, **29**, 654–655.

Chiu,J.K.H. and Chen,Y.-P.P. (2017) A comprehensive study of RNA secondary structure alignment algorithms. *Brief. Bioinform.*, **18**, 291–305.

Höchsmann,M. *et al.* (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53–62.

Jiang,T. *et al.* (1995) Alignment of trees—an alternative to tree edit. *Theor. Comput. Sci.*, **143**, 137–148.

Novák,Á. *et al.* (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, **24**, 2403–2404.

Quadrini,M. *et al.* (2019) An algebraic language for RNA pseudoknots comparison. *BMC Bioinformatics*, **20**, 161.

Schirmer,S. *et al.* (2014) Introduction to RNA secondary structure comparison. In: Gorodkin,J. and Ruzzo,W.-L. (eds.) *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods, Volume 1097 of Methods in Molecular Biology*. Humana Press, Totowa, NJ.

Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, **3**, e213.