An algebraic representation for tree alignment of RNA pseudoknotted structures

The methods proposed in the literature for RNA comparison focus mainly on pseudoknot free structures. The comparison of pseudoknotted structures is still a challenge. In this work, we propose a new algebraic representation of RNA secondary structures based on relations among hairpins in terms of nesting, crossing, and concatenation. Such algebraic representation is obtained from a defined multiple context-free grammar, which maps any kind of RNA secondary structures into extended trees, i.e., ordered trees where internal nodes are labeled with algebraic operators and leaves are labeled with loops. These extended trees permit the definition of the RNA secondary structure comparison as a tree alignment problem.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3327v1 | CC BY 4.0 Open Access | rec: 6 Oct 2017, publ: 6 Oct 2017

An Algebraic Representation for Tree Alignment of RNA Pseudoknotted Structures

Michela Quadrini, Luca Tesei and Emanuela Merelli

School of Science and Technology, Computer Science Division, Camerino, Italy

Contact:michela.quadrini@unicam.it

Introduction

Ribonucleic acid (RNA) is a linear polymer of nucleotides arranged in a sequence referred to as a *backbone*. This sequence is made of four different types of nucleotides, known as Adenine (**A**), Guanine (**G**), Cytosine (**C**) and Uracil (**U**), and folds back on itself creating *complex shapes*, known as *secondary* and *tertiary structures*. During such process, called *folding process*, each nucleotide can interact at most with one other nucleotide establishing a hydrogen bond performing Watson-Crick base pairs (**G-C** and **A-U**) or wobble base pairs (**G-U**). In $2-$dimensions, the RNA folding process can perform many RNA secondary structures; it depends on the free energy of RNA configurations. The RNA secondary structure is composed of five basic structural elements namely *hairpins*, *internal loops*, *bulges*, *helixes* (or *stacks*) and *multi-loops* [1]. Each structural element, generated when at least one base pair is performed, is a *loop*. Therefore, secondary structures are composed of loops. If no crossing interaction among loops is present, the secondary structure is said to be *pseudoknots free*, as illustrated on the left part of Figure 1 (A), otherwise it is called *pseudoknotted*, as depicted on the right part of Figure 1 (A). Both structures are also represented on the left and right part of Figure 1 (B), respectively, taking advantage of *diagrams*. Diagrams are graphs whose vertices identify nucleotides and are represented on a horizontal line; the arcs that connect two nonconsecutive vertices correspond to base pairs.

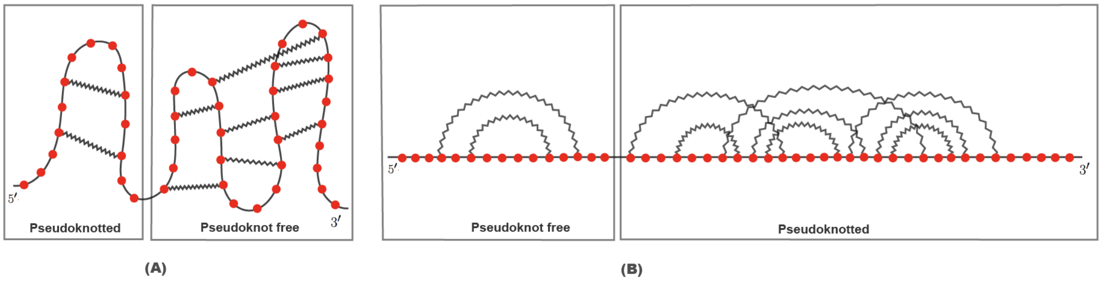

Figure 1: RNA secondary structure

RNA molecules perform a wide range of functions in biological systems, such as the catalysis of various ribozymes and the alteration of gene expression by inducing ribosomal frameshifting in many viruses. Such biological functions of RNA molecules depend on their structure. To preserve a specific function, the molecule must maintain the configuration of its secondary and tertiary structure. Therefore, structure comparison is used during the classification of RNA molecules, the prediction of the folding process and the measurement of the evolution stability. The methods proposed in literature for RNA comparison focus mainly on pseudoknot free structures. They can be generally classified into two categories: **tree edit** and **tree alignment**. The former constructs a common subtree, whereas the latter a common supertree. Tree edit is used to identify recursive structures during the folding process, while the alignment is suitable for clustering RNA molecules purely at the structure level. The comparison of pseudoknotted structures is still a challenge [2]; only a few algorithms have been developed for studying specific cases of pseudoknots due to the complexity of the problem.

In this work, we propose a new algebraic representation of RNA secondary structures that allows us to model them as an **extended tree**. The approach is based on relations among hairpins in terms of *nesting*, *crossing*, and *concatenation*. The hairpin is the basic loop of the algebraic representation.

Methods

We introduce a new algebraic representation able to model both pseudoknot free and pseudoknotted RNA secondary structures. The main novelty of this representation, with respect to other approaches present in the literature, is that we model each secondary structure as an *algebraic composition* of hairpins using a set of appropriate *operators*. The defined operators are: *concatenation* $\odot$, *nesting* $\Cap$, and *crossing* $\bowtie$. The concatenation permits to formalize that a structure is followed by another one as illustrated in Figure 2 (A). Nesting allows us to formalize the insertion of a structure into a hairpin, Figure 2 (B), while crossing operator models the crossing interactions among hairpins, as illustrated in Figure 2 (C). According to the nature of RNA molecules, nesting and interaction are well-defined if each nucleotide of the resulting structure performs at most one base pair. No restriction for concatenation is formalized since the two structures do not share nucleotides.

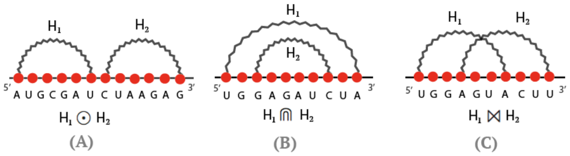
H1
H2
5′ A U G C G A U C U A A G A G 3′
H1 ⊙ H2
(A)
H1
H2
5′ U G G A G A U C U A 3′
H1 ⋒ H2
(B)
H1
H2
5′ U G G A G U A C U U 3′
H1 ⋈ H2
(C)

Figure 2: Relations among hairpins

The combination of these three operators is formalized with a multiple context-free grammar. This choice is due to the inadequacy of a Context-Free Grammar to describe the crossing dependence of pseudoknots; this can be proved by applying Ogden's Lemma [3]. The defined grammar G_{RNA} is unambiguous. In other words, the mapping from the RNA molecule to the corresponding algebraic term is a bijection. This grammar permits to model each RNA secondary structure as **extended trees**, i.e., ordered trees where internal nodes are labeled with algebraic operators and leaves are labeled with loops. The extended tree is used to formalize the pseudoknotted structure comparison as a **tree alignment problem**, which consists in determining the minimum cost over all possible alignments of the two trees. For our purposes, the cost function must be based on the edit operations, i.e. the insertion, the deletion and the substitution of unpaired nucleotides and base pairs of nucleotides, as proposed in literature as wel as the operators of the grammar G_{RNA}. If we are only interested in comparing the structure of RNAs, we may ignore the primary sequence representing all the nucleotides with the same label.

Results and Conclusions

We introduced an algebraic representation obtained from the defined multiple context-free grammar G_{RNA}, which maps any kind of RNA secondary structures into extended trees. These extended trees permit the definition of the RNA secondary structure comparison as a tree alignment problem. Moreover, this kind of trees allows us to define a measure based on tree edit distance.

We are developing a tool that permits to represent each RNA secondary structures as an extended tree and to compare the RNA structures by tree alignment. In the future, we want to extend it including also the tree edit distance.

References

1. Michael Waterman and Temple F. Smith. RNA Secondary Structure: a Complete Mathematical Analysis. Mathematical Biosciences. 1978. 42(3):257-266.

2. Stefanie Schirmer, Yann Ponty and and Robert Giegerich. Introduction to RNA Secondary Structure Comparison. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. 2014. 247-273.

3. Markus E. Nebel and Frank Weinberg. Algebraic and Combinatorial Properties of Common RNA Pseudoknot Classes with Applications. Journal of Computational Biology. 2012. 19(10):1134-1150.