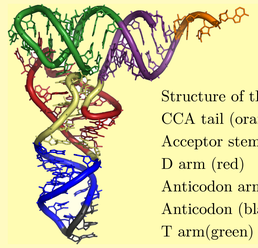
RNA Motif Discovery using Probabilistic Tree Adjoining Grammars

Patricia A. Evans and Emad Bahrami Samani {pevans,emad.b.samani}@unb.ca University of New Brunswick

Introduction

RNA is an informational molecule which plays an important role in living organisms. RNA not only is the main component in transcription in cell and protein construction but also its 3D structure allows it to be a biocatalyst. Nucleic acid targeted drug design which mainly takes advantage of the RNA in the cell is an strong hope to cure huge trouble-making diseases such as AIDS and cancer. Finding patterns in RNA structures is the first step in this way. Thus, RNA structural motif discovery has immediate applications in medicine.



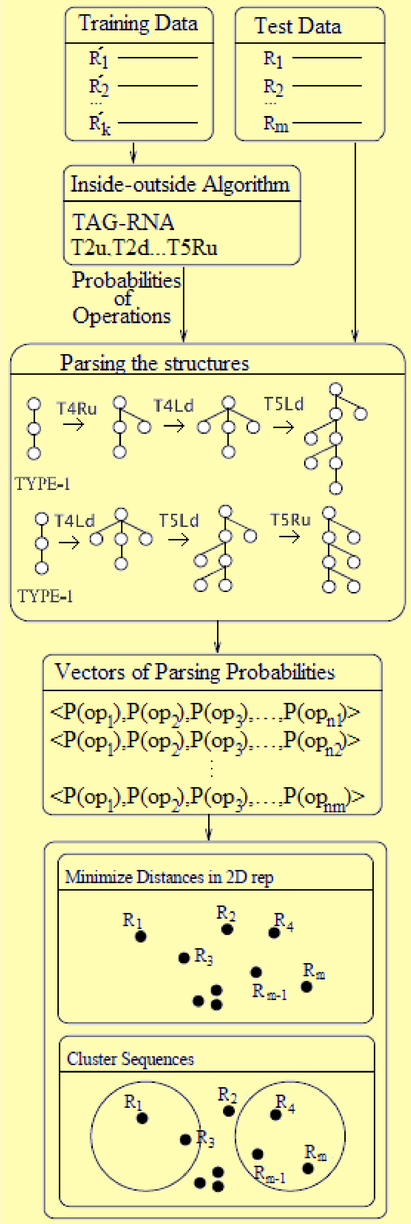
Structure of tRNA - CCA tail (orange), Acceptor stem(purple), D arm (red) Anticodon arm (blue) Anticodon (black), T arm(green)

This problem is a difficult one. Firstly because we need to deal with large amounts of biological data to recognize complex patterns. Secondly, we need methods to direct biological experiments. Pseudoknots are very important types of structure elements in RNA but it has been proven that modeling the RNA secondary structure with arbitary pseudoknots is NP-Complete. There are several methods in the literature trying to tackle this problem but a fast and accurate method that copes with pseudoknots seems necessary. This project proposes a new technique to extract the structural motifs of RNA molecules using Tree Adjoining Grammars. The main advantage of our method is that it can efficiently model pseudoknots in RNA secondary structures and extract motifs containing these structures accurately and fast. There have also been several grammatical approaches to modeling some kinds of pseudoknots. In the grammatical approaches, secondary structure modeling can be done during the process of the parsing the grammars, which can be addressed in $O(n^4)$ to $O(n^6)$ time. "Tree Adjoining Grammars" has been proposed as a useful formalism for the study of natural languages. A tree-adjoining grammar(TAG) contains two sets of elementary structures: initial trees and auxiliary trees. These elementary structures can be combined using two operations, substitution and adjunction [1].

Probabilistic TAGs

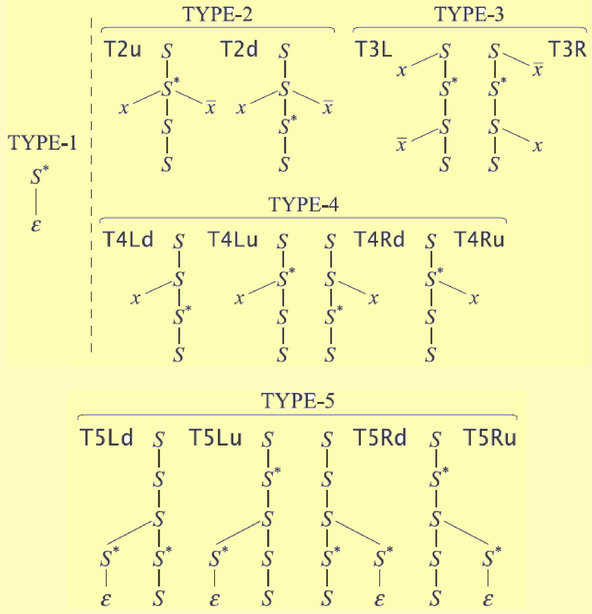
A probabilistic tree-adjoining grammar is a 5-tuple, (I, A, P_I, P_S, P_A), where I and A are Initial and Auxilary trees defined as above, P_I is a function that is known as the probability that a derivation begins with an initial tree. P_S is the probability of substitution operation and P_A denotes the probability of adjoining operations [3].

Proposed Solution in a glance



Modeling Pseudoknots

Tree adjoining grammars are described as mildly context-sensitive as they possess certain properties that make them more powerful than context-free grammars, but less powerful than context-sensitive grammars[2, 3]. Mild context sensitivity is useful for defining dependency between different parts of the string generated by the grammar. Therefore, it can be used to model pseudoknotted RNA secondary structures. In order to handle the secondary structures of RNA including pseudoknots, we will develop a novel algorithm to use the Probabilistic Tree Adjoining Grammars for RNA, denoted by TAG-RNA [2]. We will derive the TAG tree of each sequence and its annotated structure using TAG-RNA. Given an RNA sequence with its annotation of secondary structure including pseudoknots, a TAG derivation is obtained by parsing the RNA secondary structure with TAG-RNA.



Our parsing algorithm is based on the algorithm by Vijay-Shankar ([4]). TAG-RNA parser is a bottom-up parsing algorithm in nature. It uses four-dimensional dynamic programming method and can find an optimum solution with respect to some evaluation functions. The time and space complexity is $O(n^4)$, where n is the length of an input string. We use a TAG derivation process to find the common motifs between the secondary structure of two RNA [5].

Extracting Motifs

According to [3], if two points x and y, have probabilities P(x) and P(y), then their mutual information, I(x,y), is defined to be:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Mutual information compares the probability of observing x and y together with the probabilities of observing x and y independently. If there is a genuine association between x and y, then P(x,y) will be much larger than the probability P(x)P(y), and consequently $I(x, y) \gg 0$. If there is not any relationship that would be interesting for us between x and y, then P(x,y) $\simeq$ P(x)P(y), and so, $I(x, y) \simeq 0$. If x and y are in complementary distributions, then P(x,y) will be much less than P(x)P(y), forcing $I(x, y) \ll 0$.

To find novel RNA structural motifs we will calculate the mutual information between every two adjacent operation in derivation τ = (α_0, op_1(..), op_2(..),..., op_n(..)). Then we will develop a dynamic programming algorithm to find the similar patterns in these vectors. A set H of points in Euclidean space is selected so that for each sequence $s \in G$ there is a corresponding point $P(s) \in H$. Principle Coordinates Analysis (PCoA) will be used to find corresponding points in 2D space [7]. The points in H are examined to find clusters. PCoA automatically projects to the subspace where the global solution of K-means lies. RNA structural motifs are the different clusters.

References

[1] K.V. Shanker, D.J. Weir and A. K. Joshi, "Characterizing structural descriptions produced by various grammatical formalisms", *25th Annual Meeting of the Association for Computational Linguistics (ACL), 1987.*

[2] Matsui, H., et. al., "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures", *Bioinformatics, 21(11):2611-2617, 2005.*

[3] Church, K. W. and Hanks, P., "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics, 16(1):22-29, 1990.*

[4] Vijay-Shankar, k. and Joshi, A. K., "Computational Properties of Tree Adjoining Grammars", *ACL, 1985.*

[5] Uemura, Y. et. al., "Grammatically Modeling and Predicting RNA Secondary Structures", *Proc. Genome Informatics Workshop, 1995.*

[6] J. Schonfeld and D.A. Ashlock, "Evaluating Distance Measures for RNA Motif Search", *Congress on Evolutionary Computation, 2006.*

[7] *http://2008.igem.org/wiki/images/3/3f/TRNA.png*