Recent Results on Three Problems in Comparative Structural RNAomics

Shay Zakov, Nimrod Milo, Tamar Pinhas, Sivan Yogev, Erez Katzenelson, Eitan Bachmat, Yefim Dinitz, Dekel Tsur, and Michal Ziv-Ukelson*

Department of Computer Science, Ben-Gurion University of the Negev.
Email:
{zakovs,milon,tamar,yogev,erez,ebachmat,dinitz,dekelts,michaluz}@cs.bgu.ac.il

Abstract. We review our recent results on three problems in Comparative Structural RNAomics. Our contribution includes: (1) a new worst-case bound for Discrete RNA Folding, (2) Unordered Unrooted Comparisons of RNA Trees, and (3) an RNA Homology Search where the query is an RNA sequence and the output consists of Sequence-Structure homology hits allowing pseudoknots and alternative stems.
A common denominator of these three works is that they demonstrate research problems tackled within current structural RNAomics, whose solution extends to more general classical problems in Computer Science, yielding respectively: (1) a new theoretical bound for Discrete Min-Plus Matrix Multiplication, (2) a new theoretical bound for All-Pairs Cavity Bipartite Matching, and (3) new admissible heuristics to speed up Max Weighted Clique. *Source code and web-interface for the tools can be found in our website* http://www.cs.bgu.ac.il/~negevcb/.

A New Theoretical Time Bound for Discrete RNA Folding. The Classical 2D RNA Folding problem (without pseudoknots and assuming a constant bound on internal loop size) is a special form of the weighted Context Free Grammar (CFG) Parsing optimization problem. There are previous works that speed up CFG decision algorithms: Valiant's approach via Fast Boolean Matrix Multiplication, and Graham's via a boolean variant of the Four-Russians approach. There are RNA motivated extensions of these works to the weighted optimization problem: Akutsu extended Valiant's approach to yield an $O\left(n^3 \log^3 \log n / \log^2 n\right)$ time result for RNA Folding, and Frid and Gusfiled obtained an $O(n^3/\log n)$ result for Discrete RNA Folding, extending Graham's Boolean method to fast max-plus multiplication of vectors in which differences between adjacent values are confined to a small integer interval.

Recently, while studying another weighted CFG problem variant (Edit Distance with Duplications and Contractions) [3], we obtained a new bound for the discrete case of this problem. Our algorithm is an adaptation of Williams' algorithm for finite semiring matrix-vector multiplication, combined with some notions similar to the approach employed by Frid and Gusfield's algorithm. It follows the concepts of the Four-Russians approach of tabulating recurring computations. The new $O\left(n^3/\log^2 n\right)$ time algorithm

* This research was partially supported by ISF grant 478/10 and by the Frankel Center for Computer Science at Ben Gurion University of the Negev.

DOI: 10.15455/CMSR.2014.0005

applies to several problems from the domains of discrete context-free grammar parsing and RNA folding and, in particular, implies the currently asymptotically fastest algorithm for single-strand RNA folding with discrete cost functions.

Unrooted Unordered Comparison of RNA Trees. A mainstream approach to (pseudoknot free) RNA secondary structure comparison represents the structures as trees, and applies tree alignment algorithms to their comparison. Currently available bioinformatic softwares for RNA tree comparison usually apply rooted ordered tree alignment. However, there are known evolutionary events, such as segment insertions, translocations and reversals, which could be modeled as a reordering or re-rooting of branches in the corresponding trees. This motivates algorithms for Unordered Unrooted Alignment of RNA Trees.

Due to NP hardness of general Unordered Tree Edit Distance, we define a Homeomorphic Subtree Alignment variant, in which only nodes of degree 2 (or whole subtrees) can be deleted from both trees [2]. We propose and implement an efficient algorithm for this new problem variant. For the most general unrooted unordered case, the time complexity of our algorithm is $O(n^3)$, where n denotes the number of nodes in the compared trees. This improves the time complexity of previous algorithms for less general variants of the problem.

A New RNA Sequence/Structure Homology Search. A practical problem in structural RNAomics is that of genome-scale Sequence-Structure Search for conserved homologues of a given RNA *sequence*, when the structural conservation criteria are general enough to consider pseudoknots and potential dynamic alternative stem configurations.

We propose a new search engine based on a structural representation of an RNA sequence by its potential stems [1]. Potential stems in genomic sequences are identified in a preprocessing stage, and indexed. A user-provided query sequence is likewise processed, and stems from the target genomes that are similar to the query stems are retrieved from the index. Then, relevant genomic regions are identified and ranked according to the highest scoring mapping to be found between a subset of their stems versus a subset of the query stems, where the (one-to-one) stem mapping across the sets enforces conservation of cross-stem topological relations (Nested, Crossing, Adjacent, or Overlapping) within the sets. This search yields a new, NP hard, weighted variant of 2-interval pattern matching, which we solve via an efficient reduction to Max-Weighted-Clique.

References

1. N. Milo, S. Yogev, and M. Ziv-Ukelson. Stemsearch: RNA search tool based on stem identification and indexing. *Methods*, 2014.
2. N. Milo, S. Zakov, E. Katzenelson, E. Bachmat, Y. Dinitz, and M. Ziv-Ukelson. Unrooted unordered homeomorphic subtree alignment of RNA trees. *AMB*, 2013.
3. T. Pinhas, S. Zakov, D. Tsur, and M. Ziv-Ukelson. Efficient edit distance with duplications and contractions. *AMB*, 8(1):27, 2013.

DOI: 10.15455/CMSR.2014.0005