Computational Complexity and Practical Implementation of RNA Motif Search

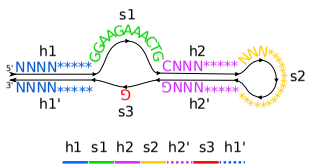
Ladislav Rampášek, Broňa Brejová

Department of Computer Science, Comenius University in Bratislava

Introduction

In this work we study the problem of RNA structural motif search, which originates in computational biology. Our work is motivated by the research of Prof. Andrej Lupták in the area of self-cleaving ribozymes, that are functional RNA molecules. Our main goal is to facilitate thorough genomic screens for these RNA structures, leading to discoveries of their novel functional occurrences in variety of organisms.

We build upon our previous work (bachelor thesis), where we proposed a new algorithm for the problem of RNA structural motif search (implemented as RNArobo). Unlike other general search tools, RNArobo can be used to find also such instances of an RNA motif that contain single letter insertions, i.e. are evolutionary more distinct. Here, our concern is the actual time performance and search results post-processing.



(a)

```
h1 s1 h2 s2 h2' s3 h1'

h1 0:0 NNNN*****:*****NNNN
h2 0:0 CNNN*****:*****NNNG

s1 1 GGAAGAAACTG
s2 0 NNN[17]
s3 0 G
```

(b)

Figure 1: *(a) A simple RNA structural motif and (b) its formal specification in form of a descriptor.*

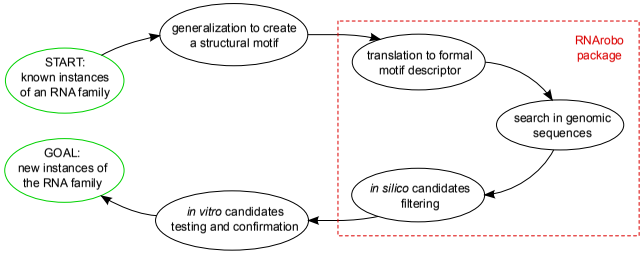


Figure 2: *Overview of the pipeline for discovering new instances of an RNA family.*

NP-completeness

Firstly, we devoted our attention to the study of computational complexity of RNA structural motif search problem (RNA-SMS). We formally define a generalized problem, the structural motif search problem (SMS), and prove it to be NP-complete. The proof is conducted by a reduction from another NP-complete problem, ONE-IN-THREE 3SAT. We subsequently show a straightforward modification of the reduction to obtain NP-completeness of RNA-SMS.

RNArobo

RNArobo is a backtracking based algorithm, and as such, it is sensitive to ordering of variables (elements). Here, we propose a data-driven method for finding a close-to-optimal element ordering. Our approach consists of two main parts: (i) heuristic proposal of possible orderings; (ii) data-driven evaluation of the proposed orderings.

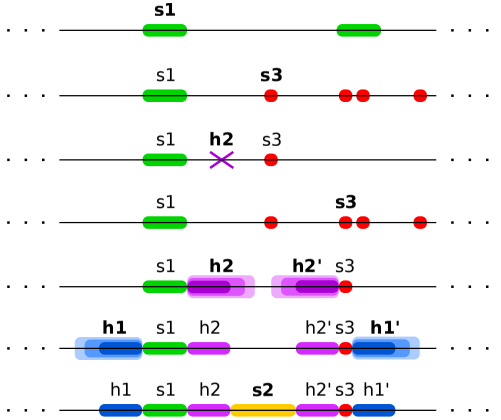


Figure 3: *An illustration of RNArobo search procedure for the motif from Figure 1. The algorithm uses a simple backtracking strategy with a fixed search ordering of elements s1, s3, h2, h1, s2. We find all matches of s1 then try to expand to an occurrence of the complete motif by recursively searching for matches of s3, h2, h1, s2 in appropriate relative positions with respect to s1. We use dynamic programming to find all matches of a particular element.*

Our experimental results demonstrate that the data-driven element ordering method implemented in RNArobo 2.0 can bring significant execution time speed-up in practice, especially in search for more complex motifs, when RNArobo 2.0 outperforms currently established tool like RNAbob and RNAmotif.

Further, in close collaboration with domain experts, we have developed tools for post-processing RNArobo search results by ranking them according to their estimated structural stability. Overall our work resulted in a practical computational subpipeline (in Figure 2 depicted as “RNArobo package”), enabling for high-throughput genome searches needed by biochemists in their pipeline for discovering new instances of an RNA family.