

# RNAmine: Frequent Stem Pattern Miner from RNAs

Michiaki Hamada<sup>1,4,6</sup>

hamada-michiaki@aist.go.jp

Koji Tsuda<sup>2</sup>

koji.tsuda@tuebingen.mpg.de

Taku Kudo<sup>3</sup>

taku@google.com

Taishin Kin<sup>4</sup>

kin-taishin@aist.go.jp

Kiyoshi Asai<sup>4,5</sup>

asai@k.u-tokyo.ac.jp

<sup>1</sup> Mizuho Information & Research Institute, Inc, 2-3 Kanda-Nishikicho, Chiyoda-ku, Tokyo, Japan

<sup>2</sup> Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

<sup>3</sup> Google Japan, Inc, 26-1 Sakuracho, Shibuya, Tokyo, Japan

<sup>4</sup> National Institute of Advanced Industrial Science and Technology (AIST), 2-43 Aomi, Koto-ku, Tokyo, Japan

<sup>5</sup> University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, Japan

<sup>6</sup> Department of Computational Intelligence and System Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, Japan

**Keywords:** functional RNA / ncRNA, stem, secondary structure, graph mining

## 1 Introduction

Finding the frequent patterns in biological sequences (DNA, RNA or amino-acid sequences) or in their structures is one of the most important challenge in bioinformatics. The frequent patterns of the sequences often have important functions, such as the regulatory elements in DNA, the catalytic site of enzymes and so on. Recently, it is revealed that a number of RNAs, which are not translated into proteins, play central roles at various biological stages. Those RNAs are called functional RNAs or non-protein-coding RNAs (ncRNAs) and attracting remarkable attention. Computational and experimental screenings have predicted a number of ncRNAs, but only few of these RNAs are classified, because their functions are still unknown. It is believed that the function of a ncRNA is highly related to not only its nucleotide sequence but also its *secondary structure*. Hence, on the RNA research, it is an important task to detect structure motifs from a set of RNA sequences. We propose a graph mining approach in order to enumerate the secondary structure motifs from unaligned RNA sequences.

## 2 Method and Results

Our algorithm takes a set of unaligned RNA sequences as the inputs and produces the secondary structure motifs that appear in the input RNA sequences more than a given threshold called a *minimum support*.

### 2.1 Graph Representation of RNA Sequence

Each RNA sequence in the inputs is represented as a directed labeled graph, called *stem graph*. Firstly, let us define the stem graph (which is a directed labeled graph) whose secondary structure is known. Such an RNA is described as a set of base pairs and we call a set of successive base pairs a *stem*. In a stem graph, a vertex corresponds to a stem and an edge describes the relative position of the stems. Each edge has a label “Parallel”, “Nested” or “Pseudoknotted” according to the relative position. When the secondary structure is unknown, each node is a *stem candidate*, not a confirmed stem. The stem candidates are derived from the base pairing probability matrix calculated by the McCaskill’s algorithm [2]. The  $(i, j)$  value of this matrix represents the

probability of the  $i$ -th nucleotide and the  $j$ -th nucleotide forming a base pair. The consecutive base pairs whose probabilities are more than  $p_{min}$  are identified as a stem candidate, but those shorter than  $l_{min}$  are discarded. Remark that the consistent secondary structure must be *clique* subgraph in the stem graph.

## 2.2 Label Taxonomy for Vertex Labels

Since we have not define the labels of vertexes yet, a set of discrete labels are introduced in the following way. The nodes of all the stem graphs are clustered and organized as a label taxonomy. Firstly, a dendrogram is generated by the hierarchical clustering of the nodes (i.e., stem candidates) from *all* the stem graphs by using the measure of dissimilarity between stems. Then, it is sliced to the layers by the given dissimilarity thresholds, and we generate a label for each cluster in each layer. The *label taxonomy* is the resulting tree of labels, and this represents similarity of vertexes.

## 2.3 Formulation as a Graph Mining Problem

In our settings, the secondary structure motif of RNA is represented as a *clique* graph, called a *stem pattern*, where node labels are taken from arbitrary layers of the label taxonomy, and the edge label is either “Parallel”, “Nested” or “Pseudoknotted”. We also define a kind of *cost* for each stem pattern. The stem pattern  $P$  *matches* to a stem graph  $G$ , if they have the same topology and the same edge labels, and every node label in  $P$  is an ancestor of that of  $G$  in the label taxonomy. If a stem pattern finds matching subgraphs in several stem graphs, the corresponding RNA sequences share a partial common structure.

In summary our task is formulated as a pure graphmining problem as below:

**Problem 1** *Given a set of directed labeled graphs, a label taxonomy, a minimum support ( $minsup$ ) and a maximum cost ( $maxcost$ ), completely enumerate every pattern  $P$  that satisfies the following three conditions: (1)  $support(P) \geq minsup$ , (2)  $P$  is a clique and (3)  $cost(P) \leq maxcost$ .*

We solve this problem by extending the gSpan algorithm [1], which is an efficient algorithm for mining general graphs.

## 2.4 Results

In the tasks of common secondary structure prediction and local motif detection from long sequences, our method performed favorably both in accuracy and in efficiency with the state-of-the-art methods such as CMFinder [3].

## Acknowledgment

This work is partially supported by “Functional RNA Project” of New Energy and Industrial Technology Development Organization.

## References

- [1] Yan, X. and Han, J., gSpan: Graph-based substructure pattern mining, *Proc. IEEE International Conference on Data Mining*, 721, 2002.
- [2] McCaskill, J.S., The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29:1105–1119, 1990.
- [3] Yao, Z., Weinberg, Z., and Ruzzo, W. L., CMfinder—a covariance model based RNA motif finding algorithm, *Bioinformatics*, 22:445–452, 2005.