

# RNA Motif Search Using the Structure to String ( $STR^2$ ) Method

Oriel Bergig, Danny Barash, and Klara Kedem  
Department of Computer Science  
Ben-Gurion University  
84105 Beer-Sheva, Israel  
Corresponding Email: klara@cs.bgu.ac.il

## Abstract

*We present a novel approach for detecting RNA shapes in given selected genes. Aside of the traditional sequence-based search methods such as BLAST and FASTA, there is a growing interest in detecting specific RNA secondary structure domains by using effective structure-based search methods such as the RNAMotif. Towards this end, we devise a new algorithm with ideas taken from computational geometry. The method, called Structure to String ( $STR^2$ ), was initially developed to detect structural motifs in the tertiary structure of proteins. It converts an RNA secondary structure into a shape representing string of characters that capture the various structural motifs. To transform an RNA secondary structure to a string of characters, we adopt an approach used in proteomics for generating a collection of fragments. We identify a library of fragments for use in RNA secondary structure where each fragment is represented by a character. A unique feature of our method is that the fragments represent the geometry of the transitions between the secondary structure elements, such as the curve of the transition between stems and loops. Consequently, we represent the secondary structures of the query and target sequences by their corresponding character string representation and seek shape similarities by applying string matching algorithms. For the RNA folding prediction we use mfold. The method is implemented efficiently using suffix trees and other economization procedures. We show examples of its applicability on aptamer domains that are functionally important and are well predicted by mfold before the conversion to strings.*

## 1. Background

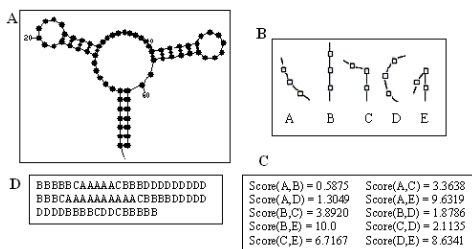
Traditionally, the search for functionally important elements in various genes has been performed using sequence similarity methods. We introduce a purely structural-based approach that works by delineating an RNA secondary

structure to a string of letters, called Structure to String ( $STR^2$ ). Initially, the  $STR^2$  method has been developed in the field of proteomics to search for structure similarities based on the geometry of the folded structures. The concept can be applied to efficiently locate shape similarities in the secondary structure of RNAs, since transitions between stems and loops and their orientation are accurately represented by the prescribed string letters.

The proposed method works as follows. Given a query structure to search among a set of target structures,  $STR^2$  will find all sub-structures of the query similar to sub-structures of the target. In our example, the query structure is the G-box [1] depicted in Figure 1(A), which is 68 nt long. We search for the G-box by predicting a set of secondary structures, namely the target structures, applying the following procedure: first, we crop from the sequence segments of 68 nt, stepping 4 nt between overlapping windows. Second, we predict the secondary structure in each window using mfold [2]. Then, we perform string matching.

The  $STR^2$  structural search method transforms the problem of structure similarity to inexact string matching [3]. It then applies fast string algorithms to solve the latter. The translation is performed using a fragment library, which consists of a small number of short fragments, each associated with a character. The  $STR^2$  method forms an alphabet with character letters (corresponding to fragments) representing each short shape segment that the target and query are composed of. In the three-dimensional case, when dealing with protein tertiary structures, this library is constructed using a simulated annealing K-means clustering method as in [4]. For RNA secondary structure we chose a library of five representative fragments (depicted in Figure 1(B)). To translate a secondary structure to a character string, we decompose the secondary structure to overlapping small fragments and translate each to a letter according to the library. Each fragment is superimposed to all five fragments in the library to pick the one with the minimal RMS distance. We then represent this fragment with the letter in the library associated with the nearest (smallest RMS

distance) library fragment. Repeating this process consecutively on the secondary structure from beginning to end, we get a shape-representing character string. We define a distance score between the characters. The score is decided according to the minimum RMS distance between the fragments (see Figure 1(C)) associated with the characters, normalized to the range from 0 to 10. This enables to calculate a similarity between two character strings, defined as the sum of the scores between any two aligned characters. A similarity between two shape-representing strings corresponds to a similarity between their correlated structures.

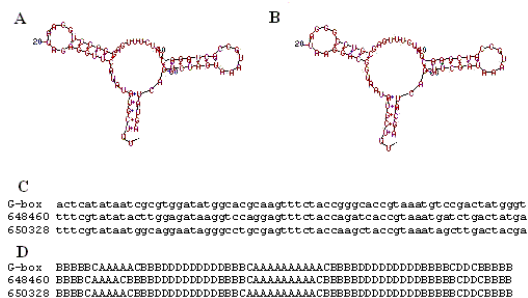


**Figure 1. The shape representing string for G-box [1] secondary structures. (A) The G-box reduced secondary structure using mfold [2]. (B) The fragment library with 5 characters, each representing 3 consecutive nucleotides. (C) The similarity score between the various characters within the fragment library. (D) G-box translated to a shape-representing string starting from the 5' end, based on the fragment library.**

## 2. Application

We first apply the Structure to String ( $STR^2$ ) method to search for the G-box (guanine binding) domain [1] of purine riboswitch in prokaryotes. Reviews on riboswitches are available in [5,6,7]. Our results shown in Figure 2 succeed to detect the G-box domain that was found by using sequence conservation in [1]. More details about the method and its validation will be given elsewhere. Furthermore, we are applying the ( $STR^2$ ) method to search for the G-box domain of purine riboswitch in eukaryotic genes. This particular task on the purine riboswitch with its G-box has not yet been successful to identify potential riboswitch sequence candidates when using methods that mostly rely on sequence based searches, albeit the success of these searches to locate a few TPP-riboswitch candidates in eukaryotes [8]. We suggest that the ( $STR^2$ ) method, which provides a novel concept not available in SequenceSniffer

[8] nor in RNA-Pattern [7] nor in the structure based similarity package RNAMotif [9] and other search methods, can be used in conjunction with the above works or by itself.



**Figure 2. (A,B) The folded secondary structures of the two G-box domains in *Bacillus halodurans* sequence target [1]. (C) The corresponding nucleotide sequences. (D) The corresponding strings that represent the structures.**

## References

- [1] M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, and R.R. Breaker, "Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria," *Cell*, Vol. 113, pp. 577-586, 2003.
- [2] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res*, Vol. 31, pp. 3406-3415, 2003.
- [3] D. Gusfield, "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology." Cambridge University Press, 1997.
- [4] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt M, "Small libraries of protein fragments model native protein structures accurately," *J Mol Biol*, Vol. 323, pp. 297-303, 2002.
- [5] W.C. Winkler, and R.R. Breaker, "Genetic control by metabolite-binding riboswitches," *Chembiochem*, Vol.4, pp.1024-1032, 2003.
- [6] E. Nudler and A.S. Mironov, "The riboswitch control of bacterial metabolism," *Trends Biochem Sci*, Vol.29, pp.11-17, 2004.
- [7] A.G. Vitreschak, D.A. Rodionov, A.A. Mironov, and M.S. Gelfand, "Riboswitches: the oldest mechanism for the regulation of gene expression?," *Trends Genet*, Vol. 20, pp.44-50, 2004.
- [8] N. Sudarsan, J.E. Barrick, and R.R. Breaker RR, "Metabolite-binding RNA domains are present in the genes of eukaryotes," *RNA* Vol.9, pp.644-647, 2003.
- [9] T.J. Macke, D.J. Ecker, R.R. Gutell, D. Gautheret, D.A. Case, and R. Sampath, "RNAMotif, an RNA secondary structure definition and search algorithm," *Nucleic Acids Res*, Vol. 29, pp.4724-4735, 2001.