Integration of RNA Search methods for Identifying Novel Riboswitch Patterns in Eukaryotes

Adaya N. Cohen[1], Klara Kedem[1], Michal Shapira[2] and Danny Barash[1]
[1]Department of Computer Science
[2]Department of Life Science
Ben-Gurion University 84105 Beer-Sheva, Israel
Corresponding Email: dbarash@cs.bgu.ac.il

Abstract

Riboswitches are RNA genetic control elements that were recently discovered in living cells. To regulate gene expression, they utilize a unique mechanism whereby small molecules bind to the aptamer or box region causing a conformational switch, without the participation of proteins. Riboswitches were initially found in the 5' UTR of bacteria, with successive discoveries in prokaryotes. Evidence for their existence in eukaryotes has prompted their scarce detection in that kingdom. Bioinformatics methods are needed in order to locate new riboswitch candidates. Several relevant search strategies have been developed and investigated, each having its own advantages and deficiencies. By merging several of these methods and integrating them in a hierarchical manner, it is possible to develop a combined strategy that will successfully locate potential candidates for the purpose of experimental validation.

1. INTRODUCTION

The focus of our current study is *riboswitches* [14, 10]. These are highly structured domains within mRNAs that precisely sense and bind metabolites, resulting in structural alterations that serve as a basis for control of gene expression. Riboswitches are typically composed of two functional domains [8]: an aptamer [5] that selectively binds its target metabolites, and an expression platform that responds to the metabolite binding and controls gene expression by allosteric means. The aptamer domain is well-conserved, whereas the expression platform can vary widely in both its sequence and secondary structure. Riboswitches have been found experimentally in prokaryotes [10, 14] and there are signs that they appear in higher organisms as well [13].

Many homologous RNAs have a common secondary structure without sharing a significant sequence similarity. Thus, searching for RNA motifs by sequence alone will likely miss important findings. There are several tools to search for RNA motifs, based on sequence and a slight structural constraint, such as the SequenceSniffer used in [13] or programs that incorporate information about numbers/lengths of stems/loops such as the RNA-Pattern used in [12]. There are more extensive search methods such as the RNAMotif [7], Fast-R [1], RNAProfile [11], RSEARCH [6], and the STR^2 search [3] developed in our group. We also plan to examine more expensive methods that utilize Hidden Markov Models (HMMs) and Stochastic Context Free Grammer (SCFG) such as RSEARCH [6] and the procedure used in [9].

2. EXPERIMENTS

The database used for the comparison is the *Bacillus halodurans* genome (NCBI accession number BA000004), tests were performed on the purine riboswitch (G-Box) known structure and members, taken from RFAM [4]. We use the G-Box domain [8] as the query sequence that consists of 67 nt.

Four types of datasets were used for examining the methods: (1) A **genomic** dataset, full genome sequence of *Bacillus halodurans* (4202353 nt); (2) A **noisy** dataset, consisting of 59 5'-UTRs of 500 nt each, taken from upstream regions of different genes. It includes 26 "noise" genes of which most encode for ribosomal proteins, and 33 genes that are involved in purine metabolism (including the reported locations taken from RFAM); (3) A dataset of **purine metabolism genes** that consists of 33 genes; (4) A **merge** dataset, RNAMotif $\bigcup$ Whiffer $\bigcup$ Fast-R outputs ("Whiffer" is a program written in our group that conceptually imitates Breaker and coworker's "Sequence Sniffer", as was described in [2, 13]).

We would like to examine the performance of each method relative to the others. There are various ways to compare, here we focus on two: **Sensitivity**, i.e. the fraction of the true matches that are actually predicted by the method, and **Specificity**, i.e. the fraction of the sequences predicted as matches that are indeed true matches. We use the following definitions for the calculations: True Positives (**TP**) are homologous findings (i.e., found in RFAM database) and are considered hits; False Positives (**FP**) are hits that are non-homologous; False Negatives (**FN**) are homologous, but not hits. We can now define Sensitivity = TP/(TP+FN), and Specificity which is the Positive Predictive Value, PPV = TP/(TP+FP). One often combines both these measures into *ROC curves*, but because of the small size of the experiments they are less likely to serve as good indicators in our case. The comparison results is shown in Table 1, where each method uses a different input as follows: **RNAMotif** uses a descriptor file based on the structure of the query; **Whiffer** uses the query as described in [8]; **Fast-R** uses seeds taken from RFAM, each of length $\approx$ 100 nt; **STR^2** uses sequence of the query; and **RNAProfile** is used in iteration mode on dataset 3 and in "all versus all" mode on dataset 4, candidates considered with fitness$\geq$0. Both RNAMotif and Whiffer fail to find the complementary matches on the genomic dataset. However, when applying the methods on datasets 2 and 3 where complementary sequences are available, both succeed to locate the missing match.

Method	Dataset	Sens.	Spec.	hits/TP/FP
RNAMotif	1	0.4	0.285	7/ 2/ 5
RNAMotif	2	0.6	1	3/ 3/ 0
RNAMotif	3	0.6	1	3/ 3/ 0
Fast-R	1	1	0.71	7/ 5/ 2
Whiffer	1	0.8	1	4/ 4/ 0
Whiffer	2	1	1	5/ 5/ 0
Whiffer	3	1	1	5/ 5/ 0
STR^2	2	1	0.416	12/ 5/ 7
STR^2	3	1	0.5	10/ 5/ 5
RNAProfile	3	0.8	0.19	21/ 4/ 17
RNAProfile	4	0.6	1	3/ 3/ 0

Table 1. Methods Comparison

3. CONCLUSIONS AND FUTURE WORK

As a consequence of our experiments, we intend to use the following procedures in eukaryotes: (1) Hierarchical global search, starting from the **merge** dataset obtained by applying FastR, RNAMotif, Whiffer on complete genomes to be followed by RNAProfile. Methods such as RSEARCH that make use of HMM and SCFG will also be explored; (2) Focused search, starting from the **genes** dataset and applying STR^2 together with other constraints imposed.

4. ACKNOWLEDGEMENTS

We thank the Lynn and William Frankel Center for Computer Sciences. The research was supported by a grant from the Israel USA binational foundation BSF 2003291.

References

[1] Bafna V, Zhang S: FastR: Fast database search tool for non-coding RNA*Proceedings of IEEE Computational Systems Bioinformatics (CSB) Conference*, 2004 :52-61.

[2] Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR: New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *PNAS* 2004; 101(17):6421-6.

[3] Bergig O, Barash D, Nudler E, Kedem K. STR2: A structure to string approach for locating G-box riboswitch shapes in pre-selected genes *In Silico Biology* 2004; 4(4):593-604.

[4] Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* 2003; 31(1):439-41.

[5] Hermann T, Patel DJ. Adaptive recognition by nucleic acid aptamers. *Science*. 2000; 287(5454):820-5.

[6] Klein RJ, Eddy SR: RSEARCH: Finding homologs of single structured RNA sequences.*BMC Bioinformatics*. 2003; 4(1):44.

[7] Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* 2001; 29(22):4724-35.

[8] Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. Riboswitches control Fundamental Biochemical Pathways in Bacillus subtilis and other Bacteria. *Cell*. 2003; 113:577-86.

[9] Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*. 2004; 306(5694):275-9.

[10] Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA and Nudler E. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* 2002; 111(5):747-56.

[11] Pavesi G, Mauri G, Stefani M, Pesole G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences.*Nucleic Acids Res* 2004; 32(10):3258-69.

[12] Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem.* 2003; 278:41148-59.

[13] Sudarsan N, Barrick JE, Breaker RR: Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* 2003; (6):644-7.

[14] Winkler W, Nahvi A and Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 2002; 419(6910):952-6.

[15] Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003; 31:3406-15.