Computational identification of riboswitches based on RNA conserved functional sequences and conformations

TZU-HAO CHANG,[1] HSIEN-DA HUANG,[2] LI-CHING WU,[3] CHI-TA YEH,[1] BAW-JHIUNE LIU,[4] and JORNG-TZONG HORNG[1,5]

[1]Department of Computer Science and Information Engineering, National Central University, Jhongli 320, Taiwan
[2]Department of Biological Science and Technology, Institute of Bioinformatics and Systems Biology, National Chiao-Tung University, Hsinchu 300, Taiwan
[3]Institute of Systems Biology and Bioinformatics, National Central University, Jhongli 320, Taiwan
[4]Department of Computer Science and Engineering, Yuan Ze University, Jhongli 320, Taiwan
[5]Department of Bioinformatics, Asia University, Wufeng 413, Taiwan

ABSTRACT

Riboswitches are *cis*-acting genetic regulatory elements within a specific mRNA that can regulate both transcription and translation by interacting with their corresponding metabolites. Recently, an increasing number of riboswitches have been identified in different species and investigated for their roles in regulatory functions. Both the sequence contexts and structural conformations are important characteristics of riboswitches. None of the previously developed tools, such as covariance models (CMs), Riboswitch finder, and RibEx, provide a web server for efficiently searching homologous instances of known riboswitches or considers two crucial characteristics of each riboswitch, such as the structural conformations and sequence contexts of functional regions. Therefore, we developed a systematic method for identifying 12 kinds of riboswitches. The method is implemented and provided as a web server, RiboSW, to efficiently and conveniently identify riboswitches within messenger RNA sequences. The predictive accuracy of the proposed method is comparable with other previous tools. The efficiency of the proposed method for identifying riboswitches was improved in order to achieve a reasonable computational time required for the prediction, which makes it possible to have an accurate and convenient web server for biologists to obtain the results of their analysis of a given mRNA sequence. RiboSW is now available on the web at http://RiboSW.mbc.nctu.edu.tw/.

Keywords: riboswitch; RNA secondary structure; regulatory RNA

INTRODUCTION

Regulatory RNAs play important roles in many essential biological processes, ranging from gene regulation to protein synthesis. Riboswitches are *cis*-acting genetic regulatory elements within a specific mRNA that can regulate both transcription and translation by binding to their corresponding targets (Coppins et al. 2007). For instance, in *Bacillus subtilis* and related species, it is estimated that >2.5% of all genes are regulated using riboswitches (Gilbert et al. 2006). Riboswitches were originally thought to occur only in the 5′ UTRs of genes; however, this notion is no longer tenable with the recent discovery of a thiamin pyrophosphate binding riboswitch in the 3′ UTR of a gene (Thore et al. 2006; Raschke et al. 2007; Wachter et al. 2007).

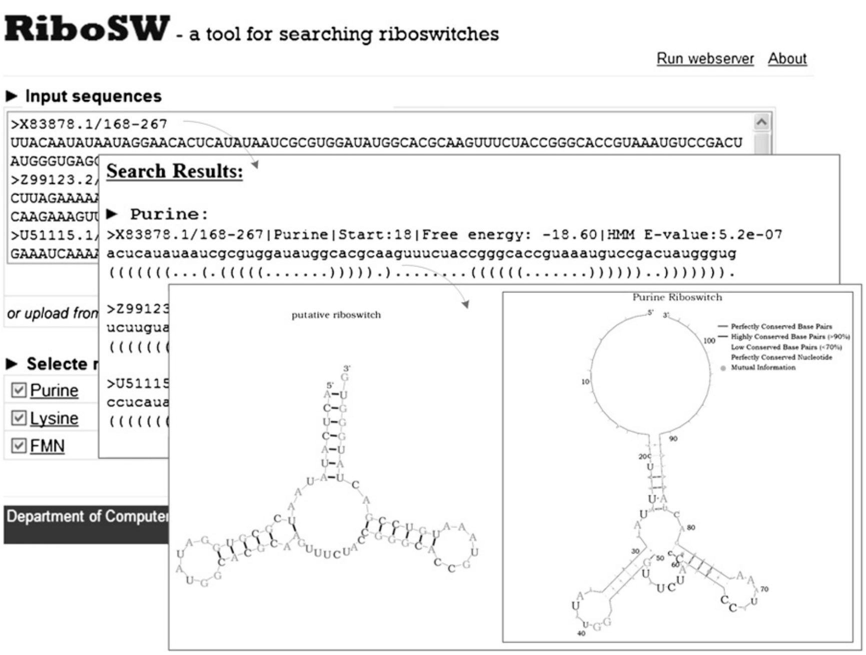
Riboswitches consist of a metabolite-responsive aptamer domain coupled with a regulatory switch. The aptamer domain forms a highly conserved specific secondary structure with the functional region, which is involved in ligand binding. Nucleotide mutation of the functional region results in a decrease in the binding affinity with the target ligand (Gilbert et al. 2007). In addition, such a nucleotide mutation also causes a change in the conformation of the secondary structure that also affects the regulatory function of the riboswitch. Therefore, a compensatory mutation within the base-paired region of the riboswitch is common because there is a need to retain the secondary structure during evolution to allow the maintenance of its functions (Fuchs et al. 2007; Wachter et al. 2007).

In recent years, three methods have been developed for the identification of riboswitches. The Rfam database

Reprint requests to: Hsien-Da Huang, Department of Biological Science and Technology, Institute of Bioinformatics and Systems Biology, National Chiao-Tung University, Hsinchu 300, Taiwan; e-mail: bryan@mail.nctu.edu.tw; fax: 886-3-5729320; or Jorng-Tzong Horng, Department of Computer Science and Information Engineering, National Central University, Jhongli 320, Taiwan; e-mail: horng@db.csie.ncu.edu.tw; fax: 886-3-4222681.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1623809.

(Griffiths-Jones et al. 2005) incorporated covariance models (CMs) when searching for riboswitches and used scoring based on a combination of the whole sequence consensus and the consensus RNA secondary structure. Riboswitch finder (Bengert and Dandekar 2004) applied a set of 13 known *B. subtilis*-like purine riboswitch sequences to establish the search program, which identifies the specific sequence elements and secondary structure in a sequence. Moreover, RibEx (Abreu-Goodger and Merino 2005) is capable of searching for 10 kinds of riboswitches against an input sequence by detecting specific sequence elements based on their sequence conservation. None of the above tools provide a web server for efficiently searching homologous instances of known riboswitches or considers two crucial characteristics of each riboswitch, such as structural conformations and sequence contexts of functional regions. Here, we developed a systematic method to identify 12 kinds of riboswitches. The method is implemented and provided as a web server, RiboSW, to efficiently and conveniently identify riboswitches within messenger RNA sequences.



RiboSW - a tool for searching riboswitches

Run webserver About

▶ Input sequences

```
>X83878.1/168-267
UUACAAUAUAAUAGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACU
AUGGGUGAG
>Z99123.2/
CUUAGAAAAA
CAAGAAAGUU
>U51115.1/
GAAAUCAAAA
```

or upload fron

▶ Selecte r

☑ Purine

☑ Lysine

☑ FMN

Department of Computer

Search Results:

▶ Purine:

```
>X83878.1/168-267|Purine|Start:18|Free energy: -18.60|HMM E-value:5.2e-07
acucauauaaucgcguggauauggcacgcaaguuucuaccgggcaccguaaauguccgacuaugggug
(((((((...(.(((((.......))))).)........((((((.......))))))..)))))))).

>Z99123
ucuugua
((((((

>U51115
ccucaua
((((((
```

putative riboswitch
Purine Riboswitch
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
5'
3'
10
20
30
40
50
60
70
80
90
100

FIGURE 1. The web interface of RiboSW.

RESULTS

The RiboSW main program is implemented using C++ programming language and runs under the Linux operating system on a PC server. Figure 1 shows the web interface of RiboSW, and a tool package is also provided on the website. The results are shown as the sequence with an underlying structure bracket notation; furthermore, information on the free energy of the secondary structure and the HMM e-value of the functional region are also provided. In addition, a RNA secondary structure graph is also drawn, and this is compared with the RNALogo graph of the corresponding Rfam riboswitch family.

Table 1 shows the search results for Riboswitch finder, RibEx, and RiboSW against the Rfam riboswitch families. For instance, there are 122 purine riboswitch sequences recorded in Rfam, and Riboswitch Finder, RibEx and RiboSW found 114, 107, and 121 purine riboswitch sequences against the Rfam records, respectively. Most riboswitches recorded in Rfam are found by RiboSW, which shows that the search ability of RiboSW is comparable with the Rfam CMs. Two riboswitches, Cobalamin and TPP, are more difficult to model with our method, which detects 77% and 88% of these Rfam members. In most other cases, RiboSW detects many more Rfam members than do the other tools.

The purine model of RiboSW was used to scan for new purine riboswitches among several species of Firmicutes (Supplemental Table S1), which is one of the phyla with sequences, available in the public domain, that have significant numbers of bacterial riboswitches. Two putative purine riboswitches were identified by our method, and the accession numbers of the located sequences, start positions, species, and information on downstream genes are presented in Supplemental Table S2. As shown in Supplemental Figure S1, the secondary structure and nucleotide sequences of the functional regions of these two putative purine riboswitches are very similar to the RNALogo graph of the Rfam purine family. Moreover, the distances from these two putative purine riboswitches to their downstream genes are 88 and 160 base pairs (bp), respectively. This corresponds to the distance distribution (in which most distances are within the range of 70–180 bp) to the downstream genes obtained from the seed members of the Rfam purine family (Fig. S2; see Supplemental Materials). Furthermore, the downstream genes of these two putative purine riboswitches are both involved in purine metabolism. The first is next to the *xpt* gene encoding xanthine phosphoribosyltransferase, which is an enzyme involved in the salvage pathway of purine nucleotide biosynthesis. In *B. subtilis*, *Lactobacillys lactis, Streptococcus pyogenes,* etc., *xpt* genes are phylogenetically well conserved, and these species were also found to have purine riboswitches located in their upstream regions as recorded in the Rfam database. The second is a gene encoding a hypothetical protein, LJ1830, which contains a uracil-xanthine permease domain that should interact with substrates such as xanthine, uracil, and

TABLE 1. The number of positive search results for RiboSW compared with other tools against the Rfam database

Riboswitch types (number of Rfam records)	Riboswitch finder (Bengert and Dandekar 2004)	RibEx (Abreu-Goodger and Merino 2005)	RiboSW
Purine (122)	114	107	121
Lysine (112)	N/A	83	111
FMN (183)	N/A	183	176
TPP (496)	N/A	467	439
Glycine (217)	N/A	184	201
SAM (298)	N/A	298	297
Cobalamin (306)	N/A	301	238
yybP-ykoY (142)	N/A	67	133
ykkC-yxkD (35)	N/A	35	35
SAM alpha (32)	N/A	N/A	32
PreQ1 (63)	N/A	N/A	63
glmS (44)	N/A	44	42

There are 122 purine riboswitch sequences recorded in Rfam. Riboswitch Finder, RibEx, and RiboSW identified 114, 107, and 121 instances of purine riboswitches against Rfam records, respectively. (N/A) Not applicable.

vitamin C. However, these two putative riboswitches were not detected by CM of the Rfam purine family. Since Rfam CMs consider overall sequence conservation, frequent compensatory mutations occurring in the helical regions of the purine riboswitch family can affect the search results. The RibEx web server returned that the regions containing these two putative riboswitches are the open reading frame (ORF), and thus no purine riboswitches are reported by RibEx. RiboSW and Riboswitch finder can detect these two putative purine riboswitches because their more sophisticated approaches take conserved functional sequences and conformation of a riboswitch into consideration. Therefore, it is clear that RiboSW is helpful when identifying novel putative riboswitches. In addition, we also used the purine model to search the upstream regions of and downstream regions from genes of various COG groups within 400 bp of the gene. Supplemental Table S3 lists the results and shows that additional purine riboswitches in three COG groups were detected. The protein functions of three COG groups—COG0503, COG0519, and COG2552—are adenine/guanine phosphoribosyltransferases and related PRPP-binding proteins, GMP synthases, and permeases, respectively. These COG groups are therefore also related to purine metabolism.

DISCUSSION

RiboSW enables user to easily search for 12 kinds of riboswitches in a sequence through the web interface. It concentrates on two characteristics of riboswitches: their RNA secondary structure and sequence conservation within their functional region; and it uses these characteristics to search for putative riboswitches within a sequence. Twelve kinds of riboswitches are modeled by our method. The phenomenon of compensatory mutations within riboswitches is also considered in RiboSW. When we compared our model with other previously developed tools; namely, Rfam CMs, RibEx, and Riboswitch finder, as shown in Table 2, we found that RiboSW is superior in many cases. RiboSW provides a convenient web server for searching most types of riboswitches with good performance, and considers two crucial parts of the riboswitch: the RNA secondary structure and the functional region. Also, RiboSW provides a good graphical visualization of results and known riboswitches using RNALogo.

Two riboswitches, Cobalamin and TPP, are more difficult to model using our method due to the high structural variation found across their member sequences. For example, as shown in Supplemental Figure S3, the Cobalamin riboswitch has many optional stems; specifically, it may or may not have the P4a, P7a, P9, and P12 stems (Nahvi et al. 2004). This makes the creation of a definitive structural descriptor more difficult than with other riboswitches. Most of the undetected member sequences of the Rfam Cobalamin family are undetected because of a failure to detect the riboswitch structure, even after we have made the structure descriptor very flexible. Despite this limitation to

TABLE 2. A comparison of RiboSW with other previously developed tools

	Rfam CMs (Griffiths-Jones et al. 2005)	RibEx (Abreu-Goodger and Merino 2005)	Riboswitch finder (Bengert and Dandekar 2004)	RiboSW
Consideration of structural conformations	Yes	No	Yes	Yes
Consideration of conserved functional sequences	No	Yes	Yes	Yes
Sequence length limitation	2 kb	40 kb	3000 kb	10 kb
Software package	Yes	No	No	Yes
Ability to define new riboswitch	Yes	No	No	Yes
Number of types of riboswitches for search	12	10	1	12

our method, RiboSW still shows an excellent ability when searching for riboswitches with a more conserved structure. It is our intention to improve RiboSW in the future by enhancing the structural search program such that optional stems may be included.

MATERIALS AND METHODS

We characterized 12 kinds of riboswitches—purine, lysine, FMN, TPP, glycine, SAM, Cobalamin, yybP-ykoY, ykkC-yxkD, SAM alpha, PreQ1, and glmS—in terms of their secondary structure and functional regions based on a literature survey and the Rfam database. This information was used to generate models of these characteristics, which were used to search for putative riboswitches in a sequence. The search process is shown in Figure 2. RiboSW involves two steps during the riboswitch search process.

The structural search

A flexible RNA structural search program was developed for RiboSW that allows numerous mispairs and bulges to be present in stem regions. To describe the conformation of riboswitches, we decompose the RNA secondary structure into several small pieces of stem. Four types of structural component (Supplemental Fig. S4) can be defined, and these are used to create a RNA secondary structure. Type 1 and type 4 structural components are defined as the first components for describing whether the whole structure is closed within a stem. Type 2 and type 3 structural components are used to describe the hairpin structure and internal stem, respectively. Most RNA secondary structures can be decomposed and described by our method (Supplemental Fig. S5) except the pseudoknot structure. For example, the purine riboswitch is made up of a three-way junction of stems as depicted in Supplemental Figure S6 using RNALogo (Chang et al. 2008). When searching for a structure similar to that of a purine riboswitch in a sequence, the structural search program looks for three structural components (Supplemental Fig. S5-a) that correspond to the structure descriptor (Supplemental Fig. S7) and that have been determined in our model at an earlier stage. Since sequence conservation is not considered during this step, compensatory mutations are not relevant to detection of a riboswitch at this point in the search. If all required structural components are discovered in the sequence region and the free energy of comprised structure is less than zero, the sequence region is then dispatched to the next step for detecting the sequence conservation of functional region.

Functional region detection

RiboSW incorporates HMMER (Eddy 1998), a well-known package for biological sequence analysis, for modeling and detecting the functional regions of riboswitches (Supplemental Fig. S8). The detected sequence regions with E-values < 10 (default cutoff in HMMER) are regarded as functional regions of the riboswitch model. For example, as shown in the RNALogo graph of the purine riboswitch (Supplemental Fig. S6), the sequences in the multibranch loop region and hairpin loop regions are more conserved than those in other regions; this is because the multibranch loop region is the high-affinity ligand binding region and thus is sensitive to single point mutations (Gilbert et al. 2007). The two hairpin loop regions are predicted to form a pseudoknot, which is an important part of the structure of the purine riboswitch (Gilbert et al. 2006). The sequence conservation within these regions is important (Supplemental Fig. S8a), and we used HMMER to model and detect the functional region. After these two steps in the search for riboswitches are completed, only results for sequences that can form the correct secondary structure with a corresponding functional region are reported.

Supplemental Figure S9 gives the comparison of the search time of INFERNAL (Rfam CMs), RibEx, Riboswitch finder, and RiboSW to search for a purine riboswitch against different sizes of input sequence (40 kb, 205 kb, 583 kb, 2754 kb, and 4197 kb). The search time required by RiboSW increases linearly with the sequence length and is slightly faster than INFERNAL. However, the performance of INFERNAL and RiboSW are comparable for long sequences.

Implementation of the web server

After the structural search and functional region detection, the putative riboswitches are shown as a sequence with bracket notation. For graphical visualization comparison with known riboswitches, RiboSW draws the RNA secondary structure of the putative riboswitch and displays it on the website with the RNALogo graph of known riboswitches. RNALogo is a useful tool for observing both the sequence and structure conservation of an RNA structural sequence. To improve the representation of different known riboswitches, we create an RNALogo graph for each riboswitch family in Rfam. In addition, the secondary structure predicted by RNAfold is also provided on the web

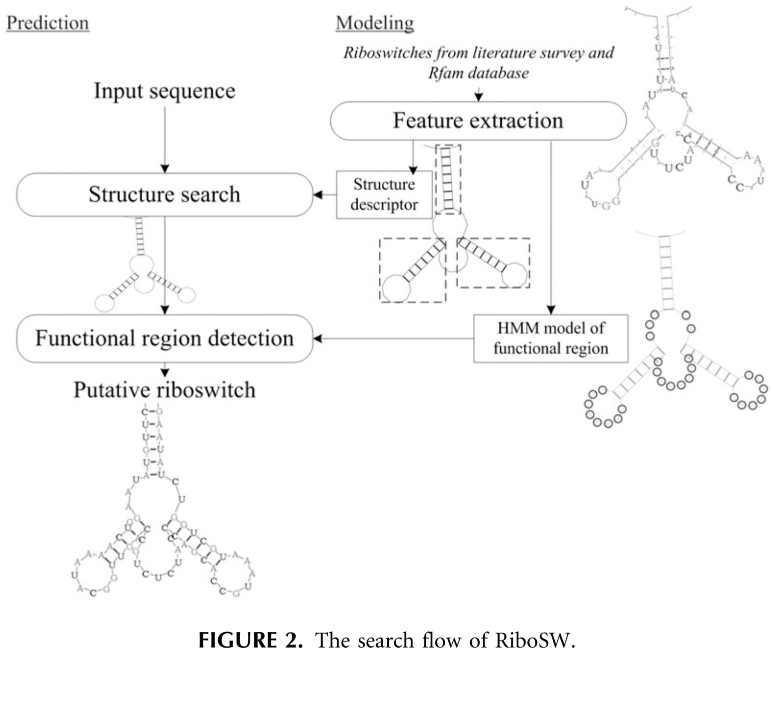

Prediction
Modeling
Riboswitches from literature survey and Rfam database
Input sequence
Feature extraction
Structure search
Structure descriptor
Functional region detection
HMM model of functional region
Putative riboswitch

FIGURE 2. The search flow of RiboSW.

page. Thus, it is very convenient to graphically compare the putative riboswitch, the minimum free energy structure, and the known riboswitch on our web server.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank the National Science Council of the Republic of China for financially supporting this research under Contract Nos. NSC 97-2627-B-009-007 and NSC 95-2311-B-009-004-MY3. Additional financial support from the National Research Program for Genomic Medicine (NRPGM), Taiwan, is gratefully acknowledged. This work was also partially supported by MOE ATU. Funding to pay for open access publication of this article was provided by the National Science Council of the Republic of China and by MOE ATU.

Received March 2, 2009; accepted April 6, 2009.

REFERENCES

Abreu-Goodger C, Merino E. 2005. RibEx: A web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res* **33:** W690–W692.

Bengert P, Dandekar T. 2004. Riboswitch finder—A tool for identification of riboswitch RNAs. *Nucleic Acids Res* **32:** W154–W159.

Chang TH, Horng JT, Huang HD. 2008. RNALogo: A new approach to display structural RNA alignment. *Nucleic Acids Res* **36:** W91–W96.

Coppins RL, Hall KB, Groisman EA. 2007. The intricate world of riboswitches. *Curr Opin Microbiol* **10:** 176–181.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Fuchs RT, Grundy FJ, Henkin TM. 2007. S-adenosylmethionine directly inhibits binding of 30S ribosomal subunits to the SMK box translational riboswitch RNA. *Proc Natl Acad Sci* **104:** 4876–4880.

Gilbert SD, Stoddard CD, Wise SJ, Batey RT. 2006. Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J Mol Biol* **359:** 754–768.

Gilbert SD, Love CE, Edwards AL, Batey RT. 2007. Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry* **46:** 13297–13309.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33:** D121–D124.

Nahvi A, Barrick JE, Breaker RR. 2004. Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res* **32:** 143–150.

Raschke M, Burkle L, Muller N, Nunes-Nesi A, Fernie AR, Arigoni D, Amrhein N, Fitzpatrick TB. 2007. Vitamin B1 biosynthesis in plants requires the essential iron sulfur cluster protein, THIC. *Proc Natl Acad Sci* **104:** 19637–19642.

Thore S, Leibundgut M, Ban N. 2006. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312:** 1208–1211.

Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. 2007. Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing of mRNAs. *Plant Cell* **19:** 3437–3450.