Bioinformatics, 32(5), 2016, 776–778
doi: 10.1093/bioinformatics/btv640
Advance Access Publication Date: 30 October 2015
Applications Note

OXFORD

Sequence analysis

Riboswitch Scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences

Sumit Mukherjee* and Supratim Sengupta

Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Mohanpur-741246, West Bengal, India

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 9, 2015; revised on October 12, 2015; accepted on October 25, 2015

Abstract

Summary: Riboswitches are non-coding RNA located in the 5′ untranslated regions where they bind a target metabolite used to specify the riboswitch class and control the expression of associated genes. Accurate identification of riboswitches is the first step towards understanding their regulatory and functional roles in the cell. In this article, we describe a new web application named Riboswitch Scanner which provides an automated pipeline for pHMM-based detection of riboswitches in partial as well as complete genomic sequences rapidly, with high sensitivity and specificity.

Availability and implementation: Riboswitch Scanner can be freely accessed on the web at http://service.iiserkol.ac.in/~riboscan/.

Contact: mukherjee.sumit89@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Riboswitches are *cis*-regulatory genomic segments that control the expression of downstream genes by undergoing conformational changes on ligand binding (Mandal and Breaker, 2004). They are mainly found in bacteria but have also been detected in a few eukaryotes (Cheah *et al.*, 2007). Riboswitches are categorized into different classes according to the type of ligand and secondary structure. High sequence level conservation and structural similarity are found in the aptamer domain of the riboswitches belonging to the same class (Nahvi *et al.*, 2002). Riboswitches control the expression of genes that are involved in biosynthesis and transport of the ligands as well as transcription factors (Singh and Sengupta, 2012). In view of their important regulatory role in bacteria and a few eukaryotes, it is essential to develop tools for accurate identification of different classes of riboswitches. Several approaches have been developed for computational identifications of riboswitches. An influential and one of the earliest programs adapted for riboswitch search is Infernal (Nawrocki *et al.*, 2009) based on the covariance model (CM). Rfam databases (Nawrocki, *et al.*, 2015) are dependent on CMs for annotation of ncRNAs. But, construction of CMs is computationally expensive and accuracy of CM is sensitive to the quality of input multiple sequence alignment (Sun and Rodionov, 2014) and no user-friendly riboswitch-specific web-servers are available for this method. RibEx (Abreu-Goodger and Merino, 2005) is a web-based tool capable of detecting riboswitches based on sequence motifs that are unique to a specific class. It is less sensitive for riboswitch classes that are characterized by short length and simple sequence motifs. The RiboSW web-server (Chang *et al.*, 2009) is able to identify 12 classes of riboswitches based on the structural conformations and sequence conservation within the functional region. But, TPP and Cobalamin riboswitches are difficult to detect with their method. Limitation of all these web-tools is that they impose upper bounds on the length of the input sequences and none of these tools are capable of detecting riboswitch from whole genome sequences. The Denison Riboswitch Detector (DRD) webserver (Havill *et al.*, 2014) was developed to predict 13 classes of riboswitches from DNA sequences based on a dynamic programming algorithm. The DRD algorithm considers mismatches which are likely to yield more false positives (Retwitzer *et al.*, 2015a). Recently, a riboswitch identification method was proposed (Retwitzer *et al.*,

2015b) that combines the accuracy of structure-based predictions with the speed of sequence-based methods by exploiting an inverse RNA folding problem solver.

In this application note, we describe a new web-based tool for accurate detection of riboswitches using a method (Singh *et al.*, 2009) based on profile Hidden Markov Models (pHMM). These are appropriate for modeling sequence profiles and searching databases for remotely homologous sequences (Eddy, 1998). By exploiting the high level of sequence conservations of the aptamer domain in every class of riboswitches, we had previously demonstrated that pHMM methods are able to identify potential riboswitches in whole genome sequences accurately and much faster than Infernal CMs search. The latest Infernal 1.1 package (Nawrocki and Eddy, 2013) currently available which uses accelerated pHMM and HMM-banded CM alignment methods and has reduced CM search times compared with older Infernal versions with little cost to detection sensitivity. However, if for searching sequences with CMs, the e-value calibration step is needed, the computing time of CM models increases substantially. Construction of riboswitch models to search of sequence databases require less computing time with pHMMs. With this motivation we developed a pHMM-based web-application named Riboswitch Scanner for predicting putative riboswitches along with their locations in genomic sequences for 24 classes of riboswitches. Moreover, users can also search new riboswitch classes. Riboswitch Scanner detects the putative riboswitches with higher accuracy than other available web-based tools for riboswitch discovery. Therefore, the Riboswitch Scanner provides more efficient alternative riboswitch detection tool that assist in the development of automated genome annotations.

2 Methods

HMM profiles of each class of riboswitches were constructed using the HMMER3 (Eddy, 2011). An HMM model was built with the *hmmbuild* program of HMMER3 packages for each class. The performances of the entire model constructed in this study were evaluated using 5-fold cross validation analysis. Sensitivity = (TP/(TP + FN)) and Specificity = (TN/(TN + FP)) are used to measure the performances of the models. TP and FN refer to true positive and false negative while TN and FP refer to true negative and false positive, respectively. RNAfold program of ViennaRNA package (Lorenz *et al.*, 2011) was used to generate the secondary structures of the detected riboswitches. All the scripts of the method were written in PHP v 5.5 and the interface was designed using HTML. The server runs on 2494 MHz Hexa core QEMU Virtual CPU version 1.4.0 under the CentOS environment. The server takes an average of 30 seconds per 1 million base pairs to generate the results.

3 Results

The webserver allows the user to submit nucleotide sequences in FASTA format. The input sequence can be a segment of a genome (such as a 5′ UTR region of a gene) or the complete genome. The user can input the sequence by pasting it in the textbox or by using the file upload facility. More than one riboswitch class can be selected at a time by checking the appropriate boxes besides each class. In order to search for riboswitches beyond the classes listed, the user can select the "other" option in the riboswitch family sections and create their own riboswitch classes by uploading at least four seed sequences in FASTA format. This generates the pHMM for the new class and searches the input sequence for instances of this user-defined riboswitch.

The results of riboswitch searches, for each of the selected riboswitch classes, in the main strand (5′–3′) and complementary strand (3′–5′) of the genome are returned in the output page in a user-friendly format. If the selected family of riboswitch is present in the genome, the riboswitch locations, *E*-value and pHMM score is generated for every detected riboswitch. Riboswitch sequences based on locations are extracted and given in the output page. Supplementary File 2 shows a sample output page. The secondary structures based on minimum free energy of riboswitches predicted by pHMM are provided using dot-bracket notations. RNAfold generated folded structure for each of the predicted riboswitches and the consensus secondary structure image (courtesy of Rfam) of the selected riboswitch family can be downloaded in a zip file from the link provided in the output page. The secondary structure images obtained using RNAfold are named in the format seq_StartPosition-EndPosition_riboswitch so that they can be easily correlated with the corresponding hits listed in the output page. Therefore, the user can further verify the pHMM-predicted riboswitches by comparing the folding pattern of the predicted riboswitches with the consensus Rfam image of this riboswitch family.

The performance of the tool for all classes was evaluated through 5-fold cross-validations. In 5-fold cross validations, each class of riboswitch seed sequence datasets was randomly divided into five sets. The training and testing of every class has been carried out five times, each time using one distinct set for training and remaining four sets for testing. The overall performance of a riboswitch model is the average performance over five sets. In comparison with other available web-based tools for riboswitch detection, Riboswitch Scanner detects riboswitches with higher sensitivity. Sensitivity and specificity of Riboswitch Scanner for all classes of riboswitches and relative comparison of sensitivity of Riboswitch Scanner with other existing web-based tools are given in the Supplementary File 1. Sensitivity and specificity are calculated based on the predictions of Rfam seed sequences.

4 Conclusions

Riboswitch Scanner was developed for predicting riboswitches with high accuracy that is complimentary to existing methods for riboswitch detection. It has several advantages over existing web-based tools for riboswitch identification: (i) It is able to scan for putative riboswitches across the largest number (24) of classes of riboswitches from partial as well as *full* genome sequences. (ii) It can predict riboswitches with higher sensitivity than any other available riboswitch-specific web-based tool. (iii) It can be easily used to search for riboswitches belonging to a new class. (iv) It provides free energy minimization-based folded structures of pHMM-predicted riboswitches for further evaluation of riboswitches based on secondary structure conservation. We believe Riboswitch Scanner will be very useful for researchers looking for an accurate riboswitch identification tool.

Acknowledgements

We thank Danny Barash, Payel Singh, Sukhen Das Mandal and Tanmoy Jana for valuable comments on the Riboswitch Scanner and Jessen T. Havill for providing us with a standalone version of their DRD package.

Funding

This work has been supported by the University Grants Commission-Israel Science Foundation (UGC-ISF) grant awarded under the India-Israel Joint Research Project 2014.

Conflict of Interest: none declared.

References

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

Chang,T.-H., *et al*. (2009) Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA*, **15**, 1426–1430.

Cheah,M.T., *et al*. (2007) Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, **447**, 497–U497.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Havill,J.T., *et al*. (2014) A new approach for detecting riboswitches in DNA sequences. *Bioinformatics*, **30**, 3012–3019.

Lorenz,R., *et al*. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26.

Mandal,M., and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, **5**, 451–463.

Nahvi,A., *et al*. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043–1049.

Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Nawrocki,E.P., *et al*. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Nawrocki,E.P., *et al*. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.

Retwitzer,M.D., *et al*. (2015a) RNAPattMatch: a web server for RNA sequence/structure motif detection based on pattern matching with flexible gaps. *Nucleic Acids Res.*, **43**, W507–W512.

Retwitzer,M.D., *et al*. (2015b) An efficient minimum free energy structure-based search method for riboswitch identification based on inverse RNA folding. *PLoS One*, **10**, e0134262.

Singh,P., *et al*. (2009) Riboswitch detection using profile Hidden Markov models. *BMC Bioinformatics*, **10**, 325.

Singh,P. and Sengupta,S. (2012) Phylogenetic analysis and comparative genomics of purine riboswitch distribution in prokaryotes. *Evol. Bioinform.*, **8**, 589–609.

Sun,E.I. and Rodionov,D.A. (2014) Computational analysis of riboswitch-based regulation. *BBA Gene. Regul. Mech*, **1839**, 900–907.