sRNATarget: a web server for prediction of bacterial sRNA targets

Yuan Cao, Yalin Zhao, Lei Cha, Xiaomin Ying, Ligui Wang, Ningsheng Shao, Wuju Li*

Beijing Institute of Basic Medical Sciences, Taiping Road 27, Haidian district, Beijing 100850, China; Wuju Li* - E-mail: liwj@nic.bmi.ac.cn; *Corresponding author

Abstract:
In bacteria, there exist some small non-coding RNAs (sRNAs) with 40–500 nucleotides in length. Most of them function as post-transcriptional regulation of gene expression through binding to their target mRNAs, in which Hfq protein acts as RNA chaperone. With the increase of identified sRNA genes in the bacterium, prediction of sRNA targets plays a more important role in determining sRNA functions. However, there are few available computational tools for predicting sRNA targets at present. Here we introduced a web server, sRNATarget, for genome-scale prediction of bacterial sRNA targets. The server is based on a recently published model which uses Naive Bayes method as the supervised method and take RNA secondary structure profile as the feature. The prediction results will be returned to the users through E-mail.

Availability: sRNATarget web server is freely available at http://ccb.bmi.ac.cn/srnatarget/.

Keywords: prediction of sRNA target; model; Naive Bayes method

Background:
Bacterial sRNAs are a class of non-coding RNAs of about 40 – 500 nucleotides in length. Most of them function as post-transcriptional regulation of gene expression through binding to their target mRNAs, in which Hfq protein acts as RNA chaperone [1]. Up to the present, even through the functions of some sRNAs have been disclosed [2], there are still many sRNAs functions to be studied. In addition, with the applications of high-throughput experimental technologies and bioinformatics methods in identifying sRNAs, more and more sRNAs have being gradually found. To speed up the processes in studying sRNA functions, developing prediction methods for sRNA targets is high demand, which provide a labor-saving way for experimental validation of sRNA targets.

Up to now, four prediction models have been presented [3-6]. In Zhang's model [3], the Smith-Waterman local sequence alignment algorithm was modified by incorporating additional information and taking ten verified sRNA-mRNA interactions as the training dataset. The prediction accuracy was 70.00%. In TargetRNA model [4], Tjaden and their colleagues developed two methods for sRNA target prediction, namely, individual base-pair method and stacked base-pair method. Through taking 12 validated sRNA-mRNA interactions as the training dataset and assuming the hybridization score of sRNA-mRNA interaction abiding by extreme value distribution, they optimized the related parameters such as the length of seed match. The prediction accuracy was about 66.67%. The third model was presented by Cossart group [5]. Four validated sRNA-mRNA interactions were used to optimize the related thermodynamic parameters, which was further applied to predict targets of their newly found sRNAs. Some predictions were verified. The fourth model, IntaRNA,

was recently presented by considering the target accessibility and seed regions. The training dataset contained 18 sRNA-mRNA interactions and the region -150 with approximately 50 around translation initiation region were taken as the potential target region. The sensitivity was 78.30% for base pairs in the training dataset. Obviously, the number of samples in each model was limited. In addition, different regions around translation initiation region were considered. To address these two problems, we have ever constructed two models, sRNATargetNB and sRNATargetSVM, using Naive Bayes method and support vector machine (SVM), respectively [7]. Both models are only for sRNAs negatively regulating their targets.

Methodology:
To construct the models for prediction of sRNA targets, we firstly collected 46 positive samples (real sRNA-mRNA interaction) and 86 negative samples (no interaction between sRNA and targets) as the training dataset. Then, the secondary structure profile of sRNA-mRNA complex was taken as the feature. The leave-one-out cross-validation (LOOCV) classification accuracy was 91.67% and 100.00% for sRNATargetNB and sRNATargetSVM, respectively. Finally, additional 22 positive samples and 1700 randomly generated negative samples were used to evaluate the performance of the models. The accuracy, sensitivity, and specificity were 93.03%, 40.90%, and 93.71% for sRNATargetNB and 80.55%, 72.73%, and 80.65% for sRNATargetSVM, respectively. In view that sRNATargetNB runs much faster and has higher specificity; it was used to construct the web server. To the best of our knowledge, there is only one web server, TargetRNA, for prediction of sRNA targets at present [4]. To facilitate the prediction of sRNA targets, we developed the web server sRNATarget. The web server, sRNATarget, was

developed using Apache, MySql, Perl, BioPerl, PHP and Javascript. It runs on HP server equipped with two 3.16G Intel Xeon CPUs and 4G memory. The interface was given in Figure 1(A). To predict sRNA targets, following three steps were suggested.

Input:
Step 1:
Choose bacterial genome:
To predict sRNA targets, the users should select bacterial genome at first. There are 727 genomes provided for the users. After the genome was chosen, the NCBI code will be automatically filled in the box next to genome box. For example, if we chose the genome for Escherichia coli K12 substr MG 1655, the code NC_000913 will be displayed.

Step 2:
Choose sRNA or enter your sRNA sequence:
After the genome was selected in step 1, all known sRNA sequences will be extracted from the related genome files, and sRNA names will be given in the sRNA list box. If users only want to predict the targets for the particular known sRNA, they can select the related sRNA name from the list box through the mouse operation. Then, the correspondent sRNA sequence will be displayed in the sRNA sequence box in FastA format. If users want to enter new sRNA sequence, they can directly enter the sequence or paste the sequence

from clipboard into the sequence box in FastA format. In **Figure 1(A)**, the sRNA dicF was demonstrated.

Step 3:
Set the score and enter your email:
To construct the server, the model sRNATargetNB was used, which was composed of 1000 classifiers. For each potential sRNA-mRNA interaction, all 1000 classifiers were used. If there are 700 classifiers to predict the interaction as the positive sample, the score will be 0.7 (i.e. 700/1000). The default score has been set to 0.5 (500/1000). In addition, the email should be provided so that the prediction results can be returned to the users.

Execution:
After step 1-3, click the submit button to do computation.

Output:
sRNATarget provided an example of genome-scale target prediction for DicF, a sRNA with 53 nt. It took about 30 minutes to finish the task. The prediction results can be accessed through the notification email (**Figure 1(B)**). There are five columns in the page, which stands for the number of entry, sRNA sequence name, mRNA sequence name, score and mRNA annotation, respectively. The prediction results are sorted in descending of scores.

Figure 1: The interface for prediction and results are displayed.

Caveat and future development:
In summary, the server sRNATarget provides a quick and labor-saving way for experimental validation of sRNA targets. To provide better support for molecular biologists, we will continue to develop models for prediction of sRNA targets in future.

References:
[1] P. Mehta *et al., Mol. Syst. Biol.* (2008) **4:** 221 [PMID: 18854820]
[2] J. Vogel & E.G.H. Wagner, *Curr. Opin. in Microbiol.* (2007) **10:** 262 [PMID: 17574901]
[3] Y. Zhang *et al., Biochem. Biophys. Res. Commun.* (2006) **343:** 950 [PMID: 16574072]
[4] B. Tjaden *et al., Nucleic Acids Res.* (2006) **34:** 2791 [PMID: 16717284]
[5] P. Mandin *et al., Nucleic Acids Res.* (2007) **35:** 962 [PMID: 17259222]
[6] A. Busch et al., *Bioinformatics* (2008) **24:**2849 [PMID:18940824]
[7] Y. Zhao *et al., Biochem. Biophys. Res. Commun.* (2008) **372:** 346 [PMID: 18501192]

Edited by P. Kangueane
Citation: Cao *et al***,** Bioinformation 3(8): 364-366 (2009)