A Hybrid Computational Approach for the Prediction of Small Non-coding RNAs from Genome Sequences

Ning Yu (1), Kyu Hong Cho (2), Qiang Cheng (1), Rafael A. Tesorero (2)
Dept. of Computer Science (1), Dept. of Microbiology (2),
Southern Illinois University Carbondale
ningyu@siu.edu, cho@micro.siu.edu, qcheng@cs.siu.edu, Rafael@siu.edu

Abstract

Researching the bacterial gene expression is a meaningful way to control and prevent the diseases which caused by bacteria. Recent researches indicate non-coding RNAs (ncRNA / sRNA) perform a variety of critical regulatory functions in bacteria. Since sRNAs have the consistent sequence characteristics, the genome-wide searching for sRNAs, especially the computational method, have become an effective way to predict the non-coding RNAs. This article proposes a hybrid computational approach for prediction of small non-coding RNAs which integrates three critical techniques, secondary structural algorithm, thermodynamic stability analysis and sequence conservation prediction. Relying on these computational techniques, our approach used to search for sRNAs in Streptococcus pyogenes which is one of the most important bacteria for human health. This search led five candidates of sRNA to be predicted as the key components of known regulatory pathways in S. pyogens.

1. Introduction

As one of most frequent pathogens of humans, the Streptococcus pyogenes are harbored by about 5-15% of normal individuals and lead to various types of suppurative infections and a very large number of diseases, such as scarlet fever, puerperal fever, and severe invasive infections etc. These infections and diseases highlight a major worldwide health concern. Efforts are directed toward the genetic mechanisms of these infections and diseases.

Bacterial small non-coding RNAs (ncRNAs / sRNAs) are regarded as novel widespread regulators of gene expression in response to environmental signals in recent researches [1,2,3]. That is, in bacteria, these RNAs or gene regulators correspond to environmental changes. Currently, in cellular processes, various roles of bacterial sRNAs have been discovered which can be categorized into two major classes based on their mode of actions. The main known sRNAs participate in post-transcriptional regulation by base-pairing with the target mRNA, which results in translation inhibition and degradation of the mRNA-sRNA complex. The other class consists of sRNAs that interact with RNA-binding proteins to modify the activity of the protein [8]. These roles of sRNAs discovered highlight the need for finding more sRNAs and identifying their functions.

An analysis of the common characteristics of these functional sRNAs emerged in 2001 which is served as a guide for systematic and genome-wide searches for sRNAs [14]. After that, several computational methods have been developed for genome-wide prediction of sRNAs. At present the annotated sRNAs in publicly available databases are mainly in Escherichia coli where more than 60 sRNAs have been identified in experiments while several more have been predicted by the computational approach [4]. Compared with E. coli, sRNAs in other bacteria were rarely discovered. And most sRNAs in other bacterial species are not so homologous as E. coli sRNAs [4]. Namely, known sRNAs in E. coli can not be employed directly to find sRNAs in other bacteria, and it limits the discovery of sRNA in other bacteria.

2. Literature Review

Ever since the first ten small non-coding RNA (ncRNA / sRNA) was serendipitously found, a number of experimental and computational methods have been proposed for finding sRNAs in the literature. Several well-known experimental methods include shotgun cloning [5], Affymetrix array [1], Northern analysis[10],
and computational ones including RNAz [6], sRNApredict [7], ISI [3], Pfold [15] etc. The majority annotated sRNAs in publicly available databases are in E. coli, however, and most of the computational methods have been tailored expressly [8]. According to the analyzed objects, the existing computational methods can be grouped into two categories - the first mainly makes use of the RNA sequences, and the second uses gene expression data. Because the gene expression levels may provide critical working information about the roles of non-coding RNAs in the ongoing biological processing at the measuring moment, the prediction of sRNAs may be more accurate and more dynamic. The gene expression testing may not be available all the times, and for this reason, the computation based on the RNA sequences has been more dominant in predicting sRNAs.

Several computational methods have used only the sequence homology, i.e., nucleotide sequence conservation among closely related species. Typical methods include sRNApredict [7], GMMI [12], etc. These methods first find the promoter and terminator signals, and then apply the matching tools to predict the sRNA sequences. Because the sequence homologous information alone uses only the primary structure of the species, which is not enough to predict the sRNAs, the accuracy might not be sufficiently high. A number of other methods have used the phylogenetic conservation properties including secondary structure conservation and thermal stability conservations. Typical methods include Pfold[15], MSARI[13], RNAz and eQRNA [6,9,10] and good prediction results have been reported. Yet another set of methods has relied on the machine learning tools such as support vector machines (SVM) [11] to predict the sRNAs from the sequences. These machine learning tools require a number of training samples, meaning that a set of known sRNAs need to be used for training the classifier, the capability of these methods applied to new species usually remain quite limited.

The above computational methods have found successful applications for E. coli RNome and very few sequenced bacterial species such as Sinorhizobium meliloti; however, the sRNAs of the majority of the bacterial RNomes still remain undiscovered. In this paper, we considered to build a computational approach that can identify the sRNAs from a user-specified set of closely related species, particularly from Streptococcus pyogenes. Also, because sRNAs may be even less than 50nt or longer than 500nt as pointed out [12], our approach handles the sRNAs between 30-550nt, though the existing methods mainly target towards sRNAs between 50-500nt [7].

In most situations where genomic promoters in bacteria are not available, our approach not only efficiently take advantage of the biological conservations of sequences, but also take advantage of the conserved secondary structures and the thermodynamic stability conservations so that the final sRNA candidates can be extracted accurately and give vital directions for experimental investigations.

3. Proposed computational approach for finding sRNAs

We built the computational approach to discover sRNAs from Streptococcus pyogenes genome sequence data. Because some computational methods have proven successful, we decided to employ them directly to avoid any repetitive work. The goal is to apply these methods to a bacterial species while the performance and accuracy can be higher. To this end, we noticed that using sequence alone (without using gene expression levels or other additional information), sequence homology and the conservation properties such as the secondary structure and thermal stability conservations, are effective in predicting sRNAs. In addition, sRNAs usually are located inside the intergenic regions (IGRs). These properties are to be utilized in constructing our computational approach. Finally, overlapping the sets of predicted sRNAs from different methods can obtain the strong sRNA candidates.

3.1. Using sequence homology

Genome homology analysis was the prerequisite of sequence conservation that requires genomic sequences of bacterial species that are appropriately diverged from the species of interest. Namely, two species must not be too similar and not be too distant. If the two species being compared are too similar, distinguishing functional sRNAs from the high background of overall sequence homology might be complex. Alternatively, the homologous functions might not be conserved if two species are too evolutionary distant.

Closely related species may have common sRNA sequences, which motivates the use of the sequence homology from several close species. In the sequences, the existence of putative sRNAs may be signified by promoters or transcription factor binding sites (TFBS). The prediction of promoter consensus sequences or
TFBS is species-specific and few of them have been experimentally determined. Therefore, many existing bioinformatics approaches used for *E. coli* cannot be directly applied to identify sRNAs to other bacterial species. However, different species may share common Rho-independent terminators. For the *Streptococcus* sequences, for instance, the promoters are unknown; yet their terminators are almost the same as those of *E. coli*. For this reason, we decided to use the terminators only in our computational approach. As some methods have been built to find the Rho-independent terminators, we directly employed one such method in our computational approach. Besides predicting the terminators, we also found the intergenic regions (IGR) using ISI [3]. The sequence homology is determined with WU-BLASTN in a way of pairwise alignments.

### 3.2. Using the secondary structure and thermal stability conservation

The secondary structure of an RNA sequence plays an important role in determining the properties of sRNAs. Figure 1 illustrates the predicted secondary structure of sRNA candidate No. 5 of *S. pyogenes* using Mfold [16]. Based on thermal stability or energy minimization, a number of computational tools are capable of finding the most likely secondary structures. RNAz [6] and eQRNA [9] are both based on the secondary structure to predict the sRNAs in yeasts. In building our computational approach, we directly employed both RNAz and eQRNA to predict sRNAs from IGR subsequences. Especially, RNAz uses the analysis based on thermodynamic stability to improve the accuracy of sRNAs prediction.

### 3.3 Relying on sequence conservation and predicted Rho-independent terminators

Accurately predicting sRNAs requires not only identifying putative transcription signals and conserved sequences but also determining their locations in the genome relative to each other and to Open Reading Frames (ORFs). Sequence conservation is suggestive of the presence of a sRNA only when found in an IGR, upstream of a putative terminator and downstream of a putative promoter (Figure 2.). Combining with the relative genomic locations of conserved sequences, transcription signals and ORFs is an effective way to overcome the limitation of pure genome-wide searching regardless of biological conservations. We employed sRNAPredict as a component of our hybrid computational approach, which is a typical method of using sequence conservations and predicted Rho-independent terminators to predict sRNAs.

### 4. Methods and Materials

The model of the hybrid computational approach used to search for putative non-coding sRNAs in *S. pyogenes* MGAS315 is illustrated in Fig.3.

#### 4.1 Extracted IGRs and Genome Homology Analysis

The genes which are annotated in the *S. pyogenes* MGAS315 were removed from the entire genome sequence of *S. pyogenes* MGAS315 for contained IGRs. We used Intergenic Sequence Inspector (ISI, [http://www.biochpharma.univ-rennes1.fr/](http://www.biochpharma.univ-rennes1.fr/)) to extract
these intergenic regions that bear conservations between phylogenetically related species.

Seven Streptococcus species were selected as the Genome Homology Analysis. The genomes were downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, December 2008), including S. equi zooepidemicus MGCS10565, S. mutans, S. suis 05ZYH33, S. sanguinis SK36, S. gordonii Challis subtr CH1, S. pneumoniae CGSP14, S. agalactiae NEM316 which were verified to have the conserved sequences in the genome homology analysis.

BLASTN program (http://blast.ncbi.nlm.nih.gov/) were used to analyze the genome homology in genome wide. The output data contain the pairwise alignments between the query sequences, that are the IGRs of S. pyogenes MGAS315, and the subject sequences, that are the genomic segments of seven homologous bacteria. The key parameters of E-value < 0.00001 and length > 30 nucleotides (nt) were kept for the BLASTN program.

4.2 Secondary Structural Conservation and Thermodynamic Stability Analysis

RNAz (http://www.tbi.univie.ac.at/~wash/RNAz) and eQRNA (ftp://selab.janelia.org/pub/software/eqrna/) which were utilized in our approach are two comparatively strict methods for secondary structural analysis compared with other methods. RNAz employs the thermodynamic stability analysis to improve the accuracy of the prediction. The multiple sequence alignments which were formatted to CLUSTALW data were used as the input source for RNAz [17]. The CLUSTALW program (http://www.ebi.ac.uk/Tools/clustalw2/index.html) employed the FASTA format as the input data. Thus, the prerequisite of executing CLUSTALW was to extract the pairwise sequences from BLASTN and generate the FASTA format. Both of RNAz and eQRNA used the window size of 150 nt and the window slide increment of 50 nt.

4.3 Sequence Conservation and Independent Terminator Analysis

ORF data, tRNA, tmRNA and rRNA were obtained from NCBI bacterial databases (ftp://ftp.ncbi.nih.gov) which include sequences, locus and names for each product. Terminators are critical for predicting sRNAs in the process of sequence conservation analysis because the promoters are usually unavailable for most bacterial species and terminators contain the biological information about the locations of ORFs, IGRs and secondary structure conservations. RNAmotif (http://cagroup.rutgers.edu/) was employed to predict Rho-independent terminators combined with TransTerm (http://transterm.cbcb.umd.edu). We selected the terminators whose probabilities are greater than 95%.

We adjusted the parameters in sRNAPredict [7], 30 bp from the start codon of an ORF and 30 bp Terminator from ORF Stop, in order to remove those which are too close to ORFs because these sequences immediately adjacent to the annotated operons or Open Reading Frames (ORFs), most likely corresponding to putative riboswitches or other cis-regulatory elements.

5. Results for Predicting sRNAs in S. pyogenes

The extracted IGRs contains 3166 sequences for both strands, totaling about 0.681MB which represents 17.9% of the full S. pyogenes MGAS315 genome. The average sequence length was 138 nucleotides (nt), with the longest one being 1501 nucleotides (nt) and the shortest one being 1 nucleotides (nt). The IGRs with a length > 12 nt were used as queries to interrogate seven streptococcus genomes with WU-BLASTN. These pairwise alignments with 2118 comparisons were filtered with the crucial parameters of E-values < 0.00001 and length > 30 nt (Table1). These pairwise alignments were to be scanned by CLUSTALW for the multiple alignments which were used by RNAz. Multiple alignments were extracted from BLASTN comparisons and the long sequences with the length > 550 nt were removed. Alignments contain two sets of strands, 5’ nucleotide strand and 3’ nucleotide strand. In order to increase the sensitivity of the model, candidates in either of the two strands were selected when any signal was identified. Overlapping eQRNA and RNAz predictions from alignments were incorporated into a single predicted RNA locus on the genome. eQRNA predicted 470 genomic segments in dual strands while RNAz predicted 187 genomic segments in dual strands. 147 genomic segments were overlapped in dual strands. And 87 sRNA candidates were obtained by merging the locus of these overlapping segments in the process of secondary structure and thermodynamic stability analysis.

To test the specificity and sensitivity of the approach, we used 86 known tRNAs and tmRNA to
verify the sRNA prediction of secondary structure and thermodynamic stability analysis. The sensitivity and specificity for the part in the prediction of the 86 known *S. pyogenes* (67 tRNA and 1 tmRNA) were assessed as Table 2. The sensitivity on known structural RNAs in our model was 65/68 (95.6%). We shuffled the sequences originated by the RNA genes and estimated the false positives if any shuffled sequence was considered as RNA. The number of false positives obtained for this part in our approach is 1/68 (1.47%).

**Table1.** Alignments number compared with *S. pyogenes* and partners after blastn.

<table>
<thead>
<tr>
<th>Query(1)</th>
<th>Subject(2)</th>
<th>Known tRNA, tmRNA</th>
<th>IGRs of <em>S. pyogenes</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. equi</em> zooepidemicus MGCS10565</td>
<td><em>S. mutans</em></td>
<td>67</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td><em>S. suis</em> 05ZYH33</td>
<td>65</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td><em>S. sanguinis</em> SK36</td>
<td>66</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td><em>S. gordonii</em> Challis substr CH1</td>
<td>66</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td><em>S. pneumoniae</em> CGSP14</td>
<td>66</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td><em>S. agalactiae</em> NEM316</td>
<td>68</td>
<td>197</td>
</tr>
</tbody>
</table>

Query(1) includes 68 tRNA and tmRNA and 3166 IGRs for both stands. Subject(2) indicates seven selected Streptococci species. Each number in the table represents the number of pairwise alignments after removing non-overlapping parts.

**Table2.** Sensitivity and specificity for the part of the secondary structure prediction and thermodynamic stability analysis.

<table>
<thead>
<tr>
<th></th>
<th>Blastn</th>
<th>eQRNA</th>
<th>RNAz</th>
<th>Overlapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real alignments</td>
<td>68/68</td>
<td>66/68</td>
<td>65/68</td>
<td>65/68</td>
</tr>
<tr>
<td>Shuffled</td>
<td>68/68</td>
<td>16/68</td>
<td>3/68</td>
<td>1/68</td>
</tr>
</tbody>
</table>

The coordinates and the orientations of 1865 ORFs, 67 tRNAs, 18 rRNAs and 1 tmRNA were marked in the genome of *S. pyogenes* MGAS315 during processing the sequence conservations. Thus, 1587 intergenic regions were extracted among ORFs and known tRNAs, rRNAs and tmRNA. RNAmotif used the motif descriptor provided by D. J. Ecker and obtained the 1778 terminators while 589 Rho-independent terminators were predicted by TransTerm with probability greater than 95%. Finally, 36 sRNA candidates were predicted by sRNApredict program and 15 sRNA candidates were left after overlapping with the set of candidates throughout eQRNA. 5 sRNA candidates were remained after final overlapping with the set of candidates throughout RNAz as the strongest candidates to encode sRNAs (Table 3).

As a result, one of five final sRNA candidates, sRNA candidate No. 3, was verified in the biological experiment [18]. The result illustrates this computational approach is an effective way to predict sRNA and the final candidates are strong enough to provide the critical directions for the further experimental verification.

**6. Conclusion**

Predicting sRNAs and finding their functions in bacterial adaptation to environment changes have the significant value in recent researches. Though the number of functional sRNAs found has a dramatic increase recently, most functional small RNAs have been unknown because experimental verifications are time-consuming and hard. Our hybrid computational approach gives vital directions to experimental verifications by integrating the analysis of sequence conservations, secondary structure and thermodynamic stability. The verification of one of sRNA candidates of *S. pyogenes* throughout experiments consolidates the reliability of our approach which will lead more sRNAs to be discovered in the future.

**7. References**


---

Table 3. Conservations of the five strongest sRNA candidates

<table>
<thead>
<tr>
<th>sRNA No.</th>
<th>Upstream ORF Name</th>
<th>Upstream Direction</th>
<th>Candidate Start</th>
<th>Candidate End</th>
<th>Candidate Direction</th>
<th>Downstream ORF Name</th>
<th>Downstream Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50S ribosomal protein L17</td>
<td>&gt;&gt;&gt;</td>
<td>80000</td>
<td>80045</td>
<td>&lt;&lt;&lt;</td>
<td>rRNA</td>
<td>&gt;&gt;&gt;</td>
</tr>
<tr>
<td>2</td>
<td>putative cell envelope proteinase</td>
<td>&gt;&gt;&gt;</td>
<td>339684</td>
<td>339787</td>
<td>&lt;&lt;&lt;</td>
<td>hypothetical protein</td>
<td>&gt;&gt;&gt;</td>
</tr>
<tr>
<td>3</td>
<td>streptolysin S associated protein</td>
<td>&gt;&gt;&gt;</td>
<td>529735</td>
<td>529922</td>
<td>&gt;&gt;&gt;</td>
<td>streptolysin S associated protein</td>
<td>&gt;&gt;&gt;</td>
</tr>
<tr>
<td>4</td>
<td>3-dehydroquinate synthase</td>
<td>&gt;&gt;&gt;</td>
<td>1285601</td>
<td>1285823</td>
<td>&gt;&gt;&gt;</td>
<td>putative acetate kinase</td>
<td>&lt;&lt;&lt;</td>
</tr>
<tr>
<td>5</td>
<td>putative aminopeptidase C</td>
<td>&lt;&lt;&lt;</td>
<td>1394707</td>
<td>1394804</td>
<td>&gt;&gt;&gt;</td>
<td>NAD synthetase</td>
<td>&lt;&lt;&lt;</td>
</tr>
</tbody>
</table>

>>> strand given in the S. pyogenes; <<< complementary strand.

---

*Figure 3.* Hybrid computational approach for predicting small non-coding RNAs in *S. pyogenes*.