Expert Systems with Applications 37 (2010) 5549–5557

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Prediction of small non-coding RNA in bacterial genomes using support vector machines

Tzu-Hao Chang [a], Li-Ching Wu [b], Jun-Hong Lin [a], Hsien-Da Huang [c], Baw-Jhiune Liu [d], Kuang-Fu Cheng [e], Jorng-Tzong Horng [a,b,f,*]

[a] *Department of Computer Science and Information Engineering, National Central University, Taiwan*
[b] *Institute of Systems Biology and Bioinformatics, National Central University, Taiwan*
[c] *Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao-Tung University, Taiwan*
[d] *Department of Computer Science and Information Engineering, Yuan Ze University, Taiwan*
[e] *Biostatistics Center and Department of Public Health, and Graduate Institute of Statistics, China Medical University, Taiwan*
[f] *Department of Bioinformatics, Asia University, Taiwan*

ARTICLE INFO

Keywords:
Expert systems
Support vector machines
Machine learning
Bioinformatics
Non-coding RNA

ABSTRACT

Small non-coding RNA genes have been shown to play important regulatory roles in a variety of cellular processes, but prediction of non-coding RNA genes is a great challenge, using either an experimental or a computational approach, due to the characteristics of sRNAs, which are that sRNAs are small in size, are not translated into proteins and show variable stability. Most known sRNAs have been identified in *Escherichia coli* and have been shown to be conserved in closely related organisms. We have developed an integrative approach that searches highly conserved intergenic regions among related bacterial genomes for combinations of characteristics that have been extracted from known *E. coli* sRNA genes. Support vector machines (SVM) were then used with these characteristics to predict novel sRNA genes.

1. Introduction

Over the past decade, RNA molecules that do not encode proteins, called functional RNAs or non-coding RNAs (ncRNAs), have been shown to play important structural and catalytic roles in the cell (Rivas, Klein, Jones, & Eddy, 2001; Storz, Opdyke, & Zhang, 2004). The bacterial ncRNAs are smaller in size that eukaryote ncRNAs, ranging from ∼50 to ∼400 nt and are termed small RNAs (sRNAs) or small regulatory RNAs (Gottesman, 2004). Small non-coding RNAs have been found to be involved in the control of: transcription, rRNA processing, RNA stability, mRNA translation, protein degradation and translocation (Wang, Ding, Meraz, & Holbrook, 2006). All of the sRNAs of *Escherichia coli* that act by base-pairing affect either the stability or translation of the mRNA target; in most cases the mRNAs are encoded in *trans* at positions on the chromosome distant from the sRNAs (Tjaden et al., 2006).

Small non-coding RNA (ncRNA) genes play critical regulatory roles in a variety of cellular processes (Wang et al., 2006), but prediction of non-coding RNA genes is a great challenge, either using an experimental or a computational approach. This is due to the characteristics of sRNAs, which include their small size, the fact that they are not translated into proteins and their variable stability. Until recently, most known sRNAs have been identified in *E. coli* and shown to be conserved in closely related organisms. In this study, we hope to use the various characteristics extracting from known sRNAs genes of *E. coli* to predict novel sRNA genes in bacteria. We developed an integrative approach for the prediction of putative sRNA genes in the related bacterial genomes using support vector machines (SVM) based on a combination of characteristics extracted from known sRNA genes.

2. Materials and methods

2.1. Genome sequence

We choose *E. coli* K12 MG1655 for our development, because *E. coli* K12 is a well-studied model organism in microbiological research. Researchers first identified and studied regulatory proteins in *E. coli* and many global analysis of gene expression have been documented for this organism (Gottesman, 2004). In addition, most known sRNAs have been identified on *E. coli*. The complete *E. coli* K12 MG1655 strain genome sequence was downloaded from EcoGene database.

2.2. Intergenic regions

The intergenic regions of *E. coli* K12 MG1655 are available from the EcoGene database. In addition, the intergenic regions of various

* Corresponding author. Address: Department of Computer Science and Information Engineering, National Central University, Taiwan.
E-mail address: horng@db.csie.ncu.edu.tw (J.-T. Horng).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.058

other bacterial genomes can be downloaded from the JCVI-CMR database. There are 2346 intergenic regions in *E. coli* K12 MG1655. Of these, 1583 (62%) intergenic regions are more than 50 nt in size. There are 3576 intergenic regions in *Salmonella enterica* serovar Typhi Ty2 and of these, 2529 (71%) intergenic regions are more than 50 nt in size.

2.3. sRNAs genes

Most known sRNAs have been identified on *E. coli*. Currently, 60 sRNAs identified in *E. coli* K12 MG1655 that are available from the EcoGene database.

2.4. System organization

The process flow of our method is depicted in Fig. 1. After we had collected genome sequences from various bacterial genomes, we identified conserved regions among the intergenic regions of related bacterial genomes. Next, we search for the existence of putative Rho-independent terminators beside the conserved regions we had found; if any Rho-independent terminator existed, we assign to this sequence the possibility that it was a candidate sRNAs, giving it a higher ranking score. Finally, we used the support vector machine model we had built to classify these sRNA candidates.

2.5. Intergenic sequences extraction

All known sRNA are encoded within intergenic regions (defined as regions between ORFs) (Wassarman, Repoila, Rosenow, Storz, & Gottesman, 2001). In addition, previous studies have indicated that no sRNAs gene resides in an intergenic region that is smaller than 50 nt and most of sRNAs genes are between 50 and 250 nt long (Hershberg, Altuvia, & Margalit, 2003). Therefore, we set 50 nt as threshold length for the selected intergenic regions during intergenic sequence extraction. We extract all intergenic regions with a length that was more than 50 nt in size and 1583 intergenic regions were pinpointed.

2.6. Finding conserved region of intergenic regions

We search for conserved regions within the intergenic regions because previous studies have indicated that small RNAs resided in intergenic regions and are generally conserved across closely related species (Gottesman, 2005; Hershberg et al., 2003; Luban & Kihara, 2007; Rivas et al., 2001). Hershberg et al. (2003) have observed conservation of sRNAs and adjacent genes among related species. Furthermore, we can clearly observe that sRNA genes are more conserved than coding genes by aligning *E. coli* with *Salmonella typhimurium* LT2, *Salmonella typhi* CT18 and *S. typhi* Ty2; sRNA genes have a higher ratio than coding-genes when sequence identity is over 85%.

Therefore, we use BLAST (Basic Local Alignment Search Tool) program to make alignment between the 1583 intergenic regions (length > 50 nt) identified in *E. coli* K12 and the 2529 intergenic regions (length > 50 nt) identified in *S. enterica* serovar *Typhi* Ty2 organism. The result shows using a bit-score as threshold of more than 80 as a filter, 809 conserved intergenic regions can be pinpointed remained between two species.

2.7. Conserved regions filtration

We download known tRNAs and rRNAs from the *E. coli* EcoGene database and create a database for querying the non-coding RNAs. After identifying the conserved interspecific intergenic regions the two enterobacterial species, we search these conserved regions by BLAST using the known tRNA/rRNA database from *E. coli* K12. If a region conserved between the two enterobacterial species was also similar to these known tRNAs or rRNAs, we remove these conserved regions from the dataset.

2.8. Building of the support vector machine model

2.8.1. Support vector machine

Support vector machine (SVM) is a supervised learning method used widely to solve classification problems. We use LibSVM, an implementation version of SVM classifier that is supplied as part of the Weka suite to perform training and prediction by the SVM approach. Weka (Waikato Environment for Knowledge Analysis)

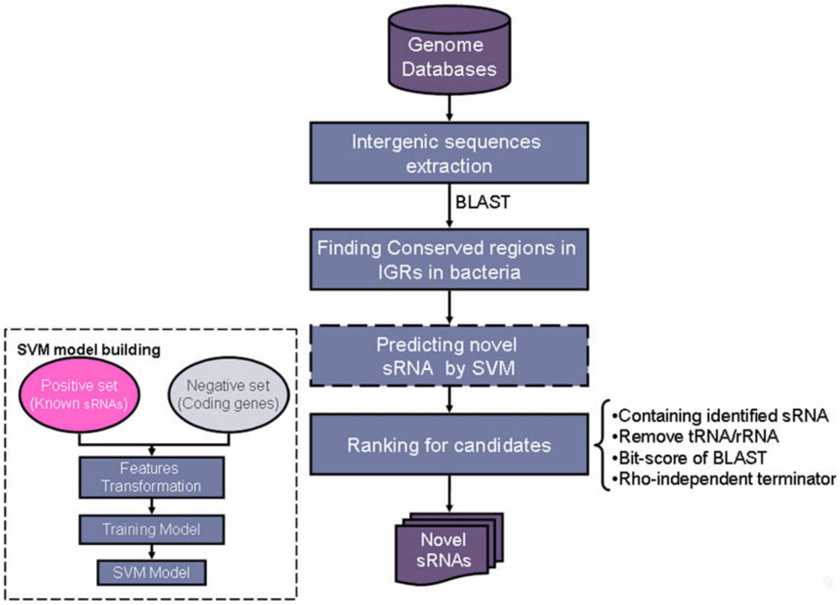
Genome Databases
Intergenic sequences extraction
BLAST
Finding Conserved regions in IGRs in bacteria
SVM model building
Positive set (Known sRNAs)
Negative set (Coding genes)
Features Transformation
Training Model
SVM Model
Predicting novel sRNA by SVM
Ranking for candidates
•Containing identified sRNA
•Remove tRNA/rRNA
•Bit-score of BLAST
•Rho-independent terminator
Novel sRNAs

Fig. 1. System organization.

is a popular suite of machine learning software designed in the Java programming language and was developed by the University of Waikato.

The set types available for the kernel function were linear function, polynomial function, radial basis function (RBF) and sigmoid function. We use the radial basis function (RBF) kernel for our approach. The two parameters in RBF kernel are γ and C. γ determines the effective range of distances between points and C determines the trade-off between margin maximization and training error minimization (Wang et al., 2006). We use a grid-search method supplied by LibSVM to identify a pair of γ and C values that gave optimal performance. The optimal parameters were $\gamma = 0.05$ and $C = 10$ and these were used in the SVM.

2.8.2. Building of training and testing set

Our positive set for the SVM training consisted of all 60 known sRNAs identified from *E. coli* K12 MG1655. The negative set for the SVM training consisted 120 coding genes randomly selected from all the coding genes of *E. coli* K12 MG1655.

2.8.3. Features transformation

Each sequence in training and testing set were transformed into a feature vector consisting of sequence composition, structural motifs, sequence conservation in related species, over-represented sequence patterns and folding minimum free energy (MFE). The sequence composition consisted of the frequency of individual nucleotide, dimers, trimers and GC content. The structural motifs consisted of the UNCG, GNRA, CUYG, AAR, CTAG motifs. The sequence conservation was the identity computed by WU-BLAST (window = 4) between the *E. coli* K12 MG1655 sequences and those in related species of the same family. These reference species were *S. typhimurium* LT2, *S. typhi* CT18 and *S. typhi* Ty2, which are members of the same enterobacterial family as *E. coli* K12. Over-represented sequence patterns are over-represented oligonucleotides in a set of sequences and these patterns were detected by the oligo-analysis tool found in regulatory sequence analysis tool (RSAT) van Helden, 2003. The minimum free energy (MFE) of each sequence was calculated using the RNAfold program (Kingsford, Ayanbule, & Salzberg, 2007).

2.8.4. Sequence composition

Sequence composition is the frequency of individual nucleotides (A, T, G, C) (4 features), of dimers (AA, AT, ..., CC) (16 features), of trimers (AAA, AAT, ..., CCC) (64 features) and the GC content (the sum of G and C frequency).

2.8.5. Structural motif

This relies on a previous study where the occurrence frequency of sequence motifs that are commonly found within RNA structural elements were identified (Carter, Dubchak, & Holbrook, 2001). We also use a set of structural motifs as features for prediction; it was hoped that not only primary sequence, but also structural level information could be useful during prediction. These structural motifs consist of the well-known sequence motifs UNCG, GNRA and CUYG (R, purine; Y, pyrimidine) found in RNA tetraloops and the AAR subsequence of the tetraloop receptor motif. In addition, the DNA sequence motif 'CTAG' (CUAG in RNA), which only occurs rarely in bacterial protein genes and non-coding regions to compared to RNA genes was included (Carter et al., 2001).

2.8.6. Sequence conservation in related species

The sequence conservation we used as a feature was the identity computed with WU-BLAST (window size = 4 with default parameters) between *E. coli* K12 MG1655 and various related species of the same family. The reference species used were *S. typhimurium* LT2, *S. typhi* CT18 and *S. typhi* Ty2, which are in the same enterobacterial family as *E. coli* K12.

Table 1
Selected features in SVM model.

Feature classes	Selected features
Dinucleotide	AT, GC, TC, TG
Trinucleotide	GAG, GTG, GCA, GGT, GAC, ATG, AGA, AAA, ACA, AAT, AGC, AAC, TGG, TAG, TCA, TGT, TTT, TAC, TTC, TCC, CAG, CGA, CAA, CAT, CTT, CCT, CCC
Structural motif	UNCG motif, CUYG motif
Sequence conservation	Conservation with Salmonella Ty2
Sequence pattern	CCCC, ACCC, AGGG, CGGC, CCAG, AACC, CGGC

2.8.7. Over-represented sequence patterns

Over-represented sequence patterns are those over-represented oligonucleotides in a set of sequences. These patterns were detected by oligo-analysis tool found in regulatory sequence analysis tool (RSAT) (van Helden, 2003).

2.8.8. Minimum free energy

The minimum free energy (MFE) of each sequence was calculated using the RNAfold program (Kingsford et al., 2007). The RNAfold program provided through the Vienna RNA package is widely used to predict possible RNA secondary structure through energy minimization. RNAfold will read an input RNA sequence and calculate its minimum free energy structure.

2.8.9. Features selection

We now had available the numerous features generated from the feature transformation step described above. However, too many features can often degrade the prediction performance of the discrimination method by over-fitting the training data (Wang et al., 2006). Therefore, we hoped to select the features from the full pool that provide a significant contribution to the prediction of sRNAs and discard the rest. We used a correlation-based feature subset selection (CFS) method for machine learning supplied by the Weka suite to select the meaningful features. After feature selection, a total of 40 features remained and these are shown in Table 1.

3. Results

3.1. Support vector machine model performance

The performance of the SVM models using the different features is shown in Table 2. The results show that our SVM model has a high accuracy when predicting sRNA sequences and therefore we used the model to identify putative sRNA sequences in conserved intergenic regions.

3.2. Validation with identified sRNAs in the ncRNAdb

The non-coding RNA database (ncRNAdb) was created as a source of information on RNA molecules that do not possess protein-coding capacity. Currently, the ncRNAdb contains >30,000 ncRNA sequences from Eukaryotes, Eubacteria and Archaea (Szymanski, Erdmann, & Barciszewski, 2007). The total number of non-coding RNAs identified from the ncRNAdb database was 437, including sRNA, tRNA, rRNA and these non-coding RNA are present in 43 species across various bacterial genomes (as shown in Table S1). Among these non-coding RNAs, there are 128 that are classified as sRNAs, including 35 identified sRNAs in *E. coli* K12 and 43 in various other bacterial genomes. These 128 sRNAs form

Table 2
The performance of our support vector machine (SVM) model.

Used features					Sn. (%)	Sp. (%)	Acc. (%)
Sequences compositions	Tetraloop motif	Sequence conservation[a]	Sequence patterns	Minimum free energy			
✓	✓	✓	✓	✓	90	100	96.6[a]
✓	✓	✓	✓	✓	85	100	95.3
✓	✓	✓	✓		85	100	95.3
✓	✓	✓			83	97	92.6
✓	✓				79	97	91.2
✓					75	97	89.9

Performance is evaluated by 10-fold cross validation.
[a] Optimized after feature selection; Sn., sensitivity; Sp., specificity; Acc., accuracy.

18 sRNAs classes made up of CsrB, CsrC, DsrA, GadY, GcvB, MicC, MicF, OxyS, RprA, RybB, RydB, RyeB, RyeE, RyfA, RyhB, SraB, SraD, SraG. We use these sRNAs as the testing set to validate whether our approach was able to detect these sRNAs correctly. The result shows that 96.9% (124/128) of "known sRNA" could be found correctly by our approach. This result indicates that our approach shows good performance when discovering known sRNAs, not only in *E. coli*, but also in other related bacterial species (Vogel & Sharma, 2005).

3.3. Validation using sRNAs candidates predicted by the PSoL tool

We use the 421 sRNA genes predicted by PSoL (Wang et al., 2006), available from the supplementary data, as a testing set to observe how many candidates can be detected by our approach. It was found that 81% of the sRNAs genes predicted by PSoL were identified by our approach. The result demonstrates our approach can reliably pinpoint putative sRNA genes and is probably at least as good as the PSoL method.

3.4. Performance comparison with sRNAFinder

For comparison with a previous study, namely sRNAFinder, we use the same dataset and criteria to evaluate our approach. According to the criteria of sRNAFinder when measuring performance, intergenic regions in which sRNAs were correctly predicted by the program were deemed true-positive predictions. There are up to 85% of known sRNAs that overlapped with conserved regions (bit-score >40) between related species. In contrast, the intergenic regions that contained no sRNAs genes in the evaluation set of known sRNAs were deemed false-positive or true-negative predictions, if the program predicted a sRNA gene in the region or not, respectively. Intergenic regions that contain sRNAs not predicted by the program were deemed false-negatives (Tjaden, 2008).

The dataset of sRNAs was made up of the 49 documented sRNAs in *E. coli* K12. Fig. 2 shows the performance of three different tools, namely QRNA, sRNAFinder and our approach. sRNAFinder gave a sensitivity of 78% and a specificity of 76%. Our approach gave a sensitivity of 84% and a specificity of 72%.

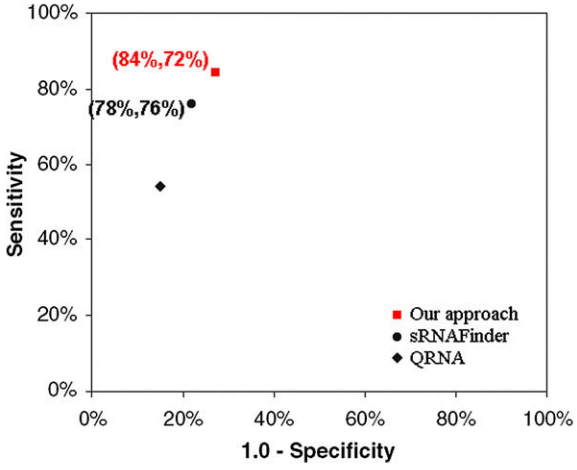

Fig. 2. A comparison of the prediction tools performance.

3.5. Criteria for selecting putative sRNAs

Although we found the conservation-based approach to be the most productive in identifying sRNA genes, a high level of conservation is not sufficient to indicate the presence of an sRNA gene (Wassarman et al., 2001). One the one hand, therefore, we selected these highly conserved regions (with high sequence similarity between related species) as our sRNA candidates. On the other hand; we also hoped to include the cases that do not show relatively high conservation between related species but did contain additional biological signals, like existence of putative or known promoters, Rho-independent terminators (an intrinsic terminator) and attenuators beside the sRNA candidate. For example, one sRNA candidate obtained only a 60 bit-score using BLAST with other related species; but this candidate has promoter and terminator in its upstream and downstream regions, respectively, this candidate is very likely to be a novel sRNA because of the presence of these biological signals. This is because promoters, Rho-independent terminators and attenuators play important roles in mechanism of regulation. The criteria we used to select sRNA candidates are listed in Table 3.

3.6. Pairwise overlap between sRNAs prediction methods

Some of the main difficulties in computational prediction of sRNA genes are the lack of benchmark data to validate the method and the difficulties associated with experimental verification on a large scale, which is expensive and time-consuming (Kulkarni & Kulkarni, 2007; Wang et al., 2006). There we apply the smart validation method suggested in PsoL. This proposes that if our results show significant agreement with other studies, this would be a validation of our method (Kulkarni & Kulkarni, 2007; Wang et al., 2006). The methods used for our comparison are listed in Table 4.

In this study, we compared our predictions to results available from previous studies. In this context, Affy is the only experimentally based method where the results are more reliable (Kulkarni & Kulkarni, 2007; Wang et al., 2006). Our method gave the largest overlap ratio with the Affy method (20%, 33/165), which suggests that our method is possibly more reliable than the other prediction methods (Tables 5 and 6). Furthermore, from the result of the pairwise comparison (shown in Table 7), it is clear that our method has the largest overlap ratio (76%) with all other sRNA prediction methods. These observations provide the strong evidence for validation of the performance of our system.

Table 3
Criteria for selecting the putative sRNAs.

Bit-score	Number of biological signals[a] found	Number of candidate in our study
⩾200	At least one	44
80–200	All	22
80–200	Two	81
50–80	All	11
40–50	All	7
Sum	–	165

[a] Biological signals: promoters, Rho-independent terminators and attenuators.

Table 4
The methods used for the pairwise comparison.

Tool	Methods	Features	References
Affy	Microarray	Microarray experiments	(Tjaden, 2008)
QRNA	HMM	Coding sequence Secondary structure conservation	(Rivas et al., 2001)
PSoL	SVM	Sequence composition Highest bits score with WU-BLAST Minimum Free Energy	(Wang et al., 2006)

Table 5
Pairwise overlap between the prediction methods and Affy.

Method (# of candidates)	QRNA (275)	PSoL (420)	Ours (165)
Affy (305)	44 (16.0%)	79 (18.8%)	33 (20.0%)

Table 6
Pairwise overlap between the various computational sRNA prediction methods.

Method (# of candidates)	QRNA (275)	PSoL (420)	Ours (165)
QRNA	–	61	47
PSoL	61	–	46
Ours	47	46	–
Sum of overlapped	108	107	93
Percentage	39	25	56

3.7. Case study I: small RNA IstR

We use our approach for the *S. enterica* subsp. *enterica* serovar Typhi Ty2 species to predict novel sRNA not found yet in other studies. We were able to pinpoint a putative sRNA by our approach. The candidate sequence is highly conserved with a documented sRNA, IstR, found in previous studies. Two sRNAs, IstR-1 and IstR-2, are encoded in the ilvB–tisAB intergenic region (Fig. S1). In some cases, IstR sRNAs can inhibits expression of downstream genes, *tisA* and *tisB*, under specific conditions (Vogel, Argaman, Wagner, & Altuvia, 2004). The candidate we found is also located in a region that has been reported as psr19 in previous study. The region psr19, a sRNA-encoding gene, was predicted to be in the intergenic region between ilvB and tisAB (Argaman et al., 2001). Furthermore, the flanking gene pairs of the IstR sRNAs gene are also conserved between *E. coli* K12 and *S. enterica* subsp. *enterica* serovar Typhi Ty2 (Fig. S2).

In addition, we wished to observe the similarity between these known and predicted IstR secondary structures only at the sequence level because similar structures often have correspondingly similar functions .We use RNALogo server (Chang, Horng, & Huang, 2008) to fold our predicted IstR sRNAs candidate and all known IstR sRNAs and RNALogo was able to report a consensus structure for all known IstR sRNAs and our predicted IstR sRNA candidate. Fig. 3 shows that the consensus structures of the predicted IstR candidate and the known IstR sRNAs, which are highly conserved.

3.8. sRNA candidates predicted by our approach

The sRNA candidates predicted by our approach are listed in Table 8. The abbreviations of the biological signals mean: KP, known promoter; PP, predicted promoter; KT, known Rho-independent terminator; PT, predicted Rho-independent terminator; PA, predicted attenuator.

4. Discussion

We have used features in addition to those used for prediction in our approach, which are mentioned above. Furthermore, in this process, we have attempted to use features that correspond to practical functions. These additional features are ones associated with the characteristics of sRNAs and include the existence of transcription factor binding sites, affinity with the Hfq protein and existence of a Rho-independent terminator sequences. However, these additional features failed to raise performance to any great extent because these additional features did not supply information that helped to predict novel sRNAs effectively. Nevertheless, we shall discuss the statistics and possible meanings of each feature and known sRNAs in the following section. During this process, the intergenic regions were scanned for promoters, for characteristic DNA sequences and for the presence a rho-independent terminator sequence (Argaman et al., 2001; Chen et al., 2002; Gottesman, 2005).

4.1. Transcription factor binding sites

We have used the existence of various distinct transcription factor binding sites and the distance between genes (sRNAs genes and protein-coding genes) as features in our SVM model for prediction. Unfortunately, it was found that no significant benefit to the success of the approach when discovering known sRNA ensued. We suggest that one reason for this might be that currently known transcription factor binding sites are too few and therefore the information cannot be used effectively to predict sRNAs.

4.2. Rho-independent terminator prediction

Rho-independent (also known as intrinsic) terminators are sequence motifs found in many prokaryotes that cause RNA transcription from DNA to stop. These termination signals typically

Table 7
Pairwise overlap between the sRNA prediction methods.

Method (# of candidates)	Affy (305)	QRNA (275)	PSoL (420)	Ours (165)	Column sum
Affy	–	44	79	33	156
QRNA	44	–	61	47	152
PSoL	79	61	–	46	186
Ours	33	47	46	–	126
Sum of overlapped	156	152	186	126	–
Percentage	51	55	44	76	

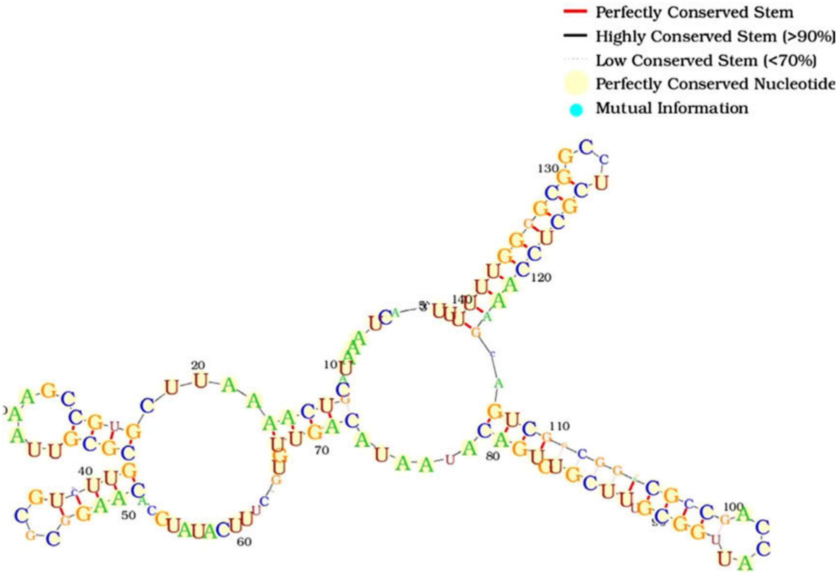
Perfectly Conserved Stem
Highly Conserved Stem (>90%)
Low Conserved Stem (<70%)
Perfectly Conserved Nucleotide
Mutual Information
5'
3'
20
40
50
60
70
80
100
110
120
130
140

Fig. 3. The consensus structure of the known IstR and our predicted IstR.

Table 8
sRNA candidates predicted by our approach.

Start position	End position	Length	Bit-score	Existence of biological signals*				
17027	17061	35	46.1		PP		PT	PA
17027	17061	35	46.1		PP		PT	PA
21079	21178	100	198			KT	PT	PA
29603	29651	49	89.7	KP	PP	KT		
160605	160755	151	236		PP		PT	
167428	167484	57	81.8		PP		PT	
190600	190857	258	341		PP		PT	
236830	237006	177	200		PP		PT	
255879	255974	96	143			KT	PT	PA
262017	262170	154	204				PT	
279347	279602	256	204		PP			
430189	430219	31	54		PP	KT	PT	PA
460961	461084	124	121	KP	PP		PT	PA
475627	475794	168	317		PP			PA
496294	496395	102	52		PP	KT	PT	PA
563945	564021	77	129	KP	PP		PT	
638732	638856	125	168				PT	PA
692642	692720	79	109		PP		PT	PA
696367	696505	139	172		PP		PT	
705247	705313	67	109	KP		KT		PA
727956	728060	105	113	KP				PA

[*] KP, known promoter; PP, predicted promoter; KT, known Rho-independent terminator; PT, predicted Rho-independent terminator; PA, predicted attenuator.

consist of a short, often GC-rich hairpin followed by a sequence enriched in thymine residues (Kingsford et al., 2007). Several previous studies have predicted novel sRNAs using terminator signal. In these circumstances, the intergenic regions were scanned for promoters and the characteristic DNA sequence and structure of a Rho-independent terminator (Argaman et al., 2001; Chen et al., 2002; Livny, Fogel, Davis, & Waldor, 2005; Rivas & Eddy, 2004; Wassarman et al., 2001; Yachie, Numata, Saito, Kanai, & Tomita, 2006).

Since candidates for novel sRNAs cannot be identified by the conventional searches used for open reading frames, we focused on transcription signals and searching for promoter sequences within a short distance upstream of a terminator (Argaman et al., 2001). Based on this terminator signals were added to our approach in the hope that performance or prediction might improve. We predict novel sRNAs in sequences where the distance between the predicted promoter and terminator were 50–400 base pairs according to a previous survey (Argaman et al., 2001). The predicted Rho-independent terminators are available from the TransTermHP database (Kingsford et al., 2007). We set confidence parameter for the putative Rho-independent terminator prediction as the default 75%. The statistics reveals there are about 50% of sRNAs in the upstream and downstream regions within 1000 nt of where we can find a putative Rho-independent terminator as predicted by TransTerm; however, when we searched all the intergenic regions, only 27.4% had sRNAs.

4.3. Attenuators

Attenuation, which involves the activation or inhibition of transcription termination at a site located between the promoter and structural genes of an operon, is a common regulatory strategy employed to sense a specific metabolic signal and enables a response that directs the RNA polymerase to either terminate transcription or transcribe the downstream genes of the operon; this system operates in many prokaryotes (Henkin & Yanofsky, 2002; Merino

Table 9
Known sRNAs with an attenuator within 500 bp of the sRNA.

sRNA name	Left position	Right position	Left position of attenuator	Right position of attenuator	Strand	Distance	Attenuator type	Attenuator regulates sRNA?
Ffs	475672	475785	475852	475895	Plus	67	Terminator	
OxyS	4156308	4156417	4156456	4156519	Plus	39	Terminator	
DsrA	2023251	2023337	2023478	2023542	Plus	141	Terminator	
			2022900	2022940	Minus	311	Terminator	
SokB	1490143	1490198	1490086	1490140	Minus	3	Terminator	Yes
SokC	16952	17006	16895	16950	Minus	2	Terminator	Yes
RttR	1286289	1286459	1285758	1285805	Minus	484	Terminator	
Tff	189712	189847	189648	189703	Minus	9	Anti-	Yes
RdlA	1268546	1268612	1269024	1269081	Plus	412	Terminator	Yes
			1268489	1268546	Minus	0	Terminator	Yes
RdlB	1269081	1269146	1269559	1269616	Plus	412	Terminator	Yes
			1269024	1269081	Minus	0	Terminator	Yes
RdlC	1269616	1269683	1269559	1269616	Minus	0	Terminator	Yes
RdlD	3698159	3698222	3698101	3698159	Minus	0	Terminator	Yes
RyeA	1921090	1921338	1921385	1921441	Plus	47	Anti–anti-	
RyeB	1921188	1921308	1921327	1921422	Plus	19	Anti-	Yes
			1921139	1921175	Minus	13	Anti-	Yes
IsrB	1985863	1986022	1986039	1986099	Plus	17	Terminator	Yes
RyeE	2165136	2165221	2165539	2165588	Plus	318	Terminator	
MicA	2812824	2812901	2812638	2812699	Minus	125	Terminator	Yes
OmrA	2974124	2974211	2974315	2974363	Plus	104	Terminator	Yes
OmrB	2974332	2974407	2974531	2974580	Plus	124	Terminator	Yes
RygD	3192745	3192887	3192932	3192976	Plus	45	Terminator	Yes
PsrO	3309247	3309420	3309429	3309492	Plus	9	Anti–anti-	
RyjA	4275950	4276089	4276266	4276329	Plus	177	Terminator	Yes
			4275535	4275591	Minus	359	Anti–anti-	Yes
IstR	3851141	3851280	3851714	3851794	Plus	434	Anti-	Yes
			3850898	3850965	Minus	176	Anti–anti-	Yes
RygE	3193121	3193262	3192932	3192976	Minus	145	Terminator	
RseX	2031673	2031763	2031501	2031546	Minus	127	Terminator	

& Yanofsky, 2005). The regulatory elements involved in attenuation are called attenuators and are cis-regulatory elements that can modulate transcription elongation or translation initiation (Gama-Castro et al., 2008).

However, exactly how attenuation and repression work together to regulate the expression of an operon is not known, but it is thought that repression provides the basic on–off switch and attenuation modulates the precise level of gene expression that occurs (Brown, 2002). The latest version of the RegulonDB database was able to provide information on predicted attenuators (Gama-Castro et al., 2008; Merino & Yanofsky, 2005). We have observed a phenomenon whereby many known sRNAs (24/60, 40%) were located beside a putative attenuator and within a very short distance of it (<500 nt); in some of these cases, the sRNA was exactly beside an attenuator (the data is shown in Table 9. Three cases are shown as Figs. S3–S5 with the hairpins depicted by the dotted line representing the putative attenuators and the arrows depicted in bold solid representing the sRNAs (Gama-Castro et al., 2008)).

We observe that sRNAs and attenuator are often located in the upstream regulatory region of operons and this phenomenon might imply that the sRNAs and attenuator both play important roles in the mechanism of genes regulation in prokaryotes.

4.4. Hfq protein

The conserved RNA-binding protein Hfq modulates the stability or the translation of mRNAs and has been shown to interact with some small regulatory RNAs (i.e. DsrA, RyhB, Spot42 RNA, OxyS) in *E. coli* that act by base-pairing (Geissmann & Touati, 2004; Moller et al., 2002; Zhang et al., 2003). Several previous studies indicate that Hfq stabilizes the small RNAs and mediates their interaction with the target mRNA by altering the target RNA structure or by interfering with ribosome binding (Aiba, 2007; Valentin-Hansen, Eriksen, & Udesen, 2004). However the precise mechanism by which Hfq regulation occurs remains unclear (Geissmann & Touati, 2004; Moller et al., 2002; Zhang et al., 2003). Hfq protein does not have a precise target sequence but appears to bind preferentially to small, single-stranded AU rich RNA segments (Moller et al., 2002; Zhang et al., 2003). Up to now, more than 30% of the known sRNAs in *E. coli* K-12 have been found to undergo Hfq-binding (Zhang et al., 2003, 2006). All the known sRNAs targets that binding Hfq are listed in Table 10.

In *E. coli*, a search for sRNAs that bind to Hfq, an RNA chaperone implicated in non-coding NA function, has yielded several novel non-coding RNAs not found by the other methods (Chen et al., 2002; Gottesman, 2005; Zhang et al., 2003). Therefore, we have developed a method to predict possible Hfq-binding sites in intergenic regions in the hope that this might help the discovery of unknown sRNAs that can bind Hfq protein and have never been found by other prediction methods. We use the RNAFold program to fold the secondary structures of known sRNAs and then search for AU rich region between two stem-loops in single strand structure, which are the criteria previous studies have suggested (Geissmann & Touati, 2004; Moller et al., 2002; Zhang et al., 2003, 2006) (three cases are shown in Fig. 4) (Gottesman, 2004). If an AU rich region between two stem-loops in single strand structure can be found, this region is considered to be a putative Hfq-binding site.

Table 10
All known small RNA targets that bind to Hfq.

Known small RNA targets			
RydC	Qrr	RyhB	DicF
OxyS	Spot 42	GcvB	MicF
RyeB	RyeE	MicC	RprA
DsrA	MicA/SraD	RybB	RybB
SraE/OmrB/RygB	SraJ	RyeF	MicA
SraH	SgrS	SroC	OmrA/RygA
GadY			

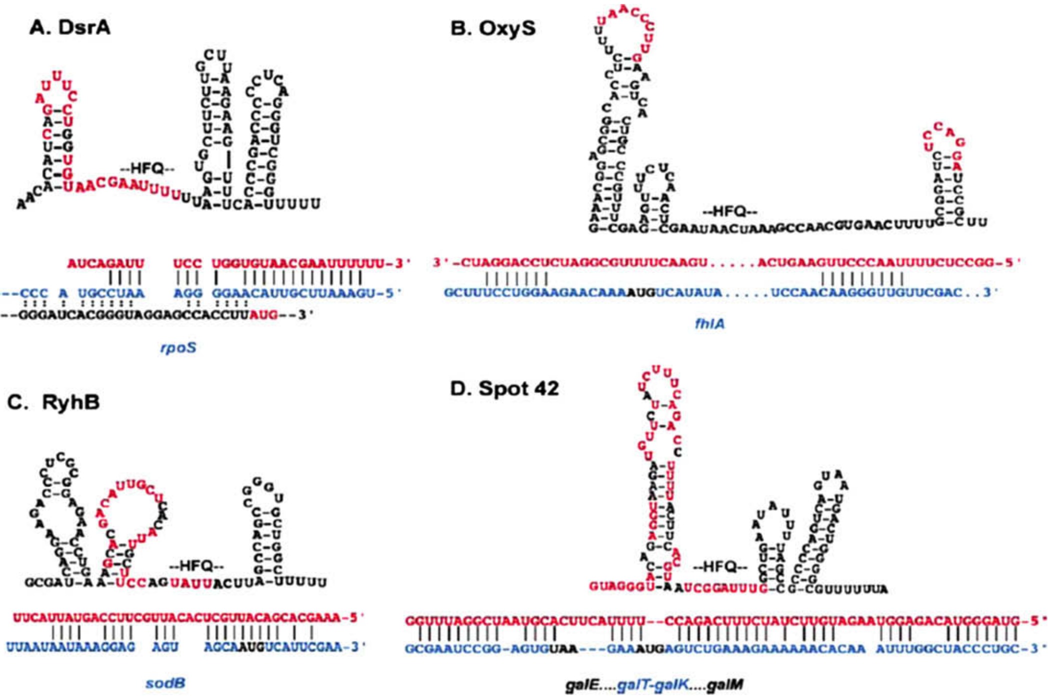
A. DsrA
--HFQ--
AUCAGAUU UCC UGGUGUAACGAAUUUUUU-3'
---CCC A UGCCUAA AGG GGAACAUUGCUUAAAGU-5'
---GGGAUCACGGGUAGGAGCCACCUUAUG--3'
rpoS
B. OxyS
--HFQ--
3'-CUAGGACCUCUAGGCGUUUUCAAGU.....ACUGAAGUUCCCAAUUUUCUCCGG-5'
GCUUUCCUGGAAGAACAAAAUGUCAUAUA.....UCCAACAAGGGUUGUUCGAC..3'
fhlA
C. RyhB
--HFQ--
UUCAUUAUGACCUUCGUUACACUCGUUACAGCACGAAA-5'
UUAAUAAUAAAGGAG AGU AGCAAUGUCAUUCGAA-3'
sodB
D. Spot 42
--HFQ--
GGUUUAGGCUAAUGCACUUCAUUUU--CCAGACUUUCUAUCUUGUAGAAUGGAGACAUGGGAUG-5'
GCGAAUCCGG-AGUGUAA---GAAAUGAGUCUGAAAGAAAAAACACAA AUUUGGCUACCCUGC-3'
galE....galT-galK....galM

Fig. 4. The four Hfq-binding sites of sRNAs, which is taken from a figure in Gottesman (2004).

In total, 15 such sRNAs have been identified in *E. coli* and we can find 10 (66.7%) of these sRNAs targets using our designed Hfq-binding sites searching method. However, we are not satisfied with the prediction performance for Hfq-binding sites. Predicting correctly Hfq-binding sites is a difficult challenge at present and we believe that a greater number of novel sRNAs will be detected using methods other than predicting Hfq-binding sites in bacterial genomes.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2010.02.058.

References

Aiba, H. (2007). Mechanism of RNA silencing by Hfq-binding small RNAs. *Current Opinion in Microbiology, 10*, 134–139.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G. H., Margalit, H., et al. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Current Biology, 11*, 941–950.

Brown, T. A. (2002). *Genomes 2*. Oxfordshire: BIOS Scientific Publishers..

Carter, R. J., Dubchak, I., & Holbrook, S. R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research, 29*, 3928–3938.

Chang, T. H., Horng, J. T., & Huang, H. D. (2008). RNALogo: A new approach to display structural RNA alignment. *Nucleic Acids Research, 36*, W91–W96.

Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J., et al. (2002). A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems, 65*, 157–177.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., et al. (2008). RegulonDB (version 6.0): Gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research, 36*, D120.

Geissmann, T. A., & Touati, D. (2004). Hfq, a new chaperoning role: Binding to messenger RNA determines access for small RNA regulator. *EMBO Journal, 23*, 396–405.

Gottesman, S. (2004). The small RNA regulators of *Escherichia coli*: Roles and mechanisms . *Annual Review of Microbiology, 58*, 303–328.

Gottesman, S. (2005). Micros for microbes: Non-coding regulatory RNAs in bacteria. *Trends in Genetics, 21*, 399–404.

Henkin, T. M., & Yanofsky, C. (2002). Regulation by transcription attenuation in bacteria: How RNA provides instructions for transcription termination/antitermination decisions. *Bioessays, 24*, 700–707.

Hershberg, R., Altuvia, S., & Margalit, H. (2003). A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Research, 31*, 1813–1820.

Kingsford, C. L., Ayanbule, K., & Salzberg, S. L. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology, 8*, R22.

Kulkarni, R. V., & Kulkarni, P. R. (2007). Computational approaches for the discovery of bacterial small RNAs. *Methods, 43*, 131–139.

Livny, J., Fogel, M. A., Davis, B. M., & Waldor, M. K. (2005). sRNAPredict: An integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research, 33*, 4096–4105.

Luban, S., & Kihara, D. (2007). Comparative genomics of small RNAs in bacterial genomes. *OMICS, 11*, 58–73.

Merino, E., & Yanofsky, C. (2005). Transcription attenuation: A highly conserved regulatory strategy used by bacteria. *Trends in Genetics, 21*, 260–264.

Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G., et al. (2002). Hfq: A bacterial Sm-like protein that mediates RNA–RNA interaction. *Molecular Cell, 9*, 23–30.

Rivas, E., & Eddy, S. R. (2004). Noncoding RNA gene detection using comparative sequence analysis. Feedback.

Rivas, E., Klein, R. J., Jones, T. A., & Eddy, S. R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current Biology, 11*, 1369–1373.

Storz, G., Opdyke, J. A., & Zhang, A. (2004). Controlling mRNA stability and translation with small, noncoding RNAs. *Current Opinion in Microbiology, 7*, 140–144.

Szymanski, M., Erdmann, V. A., & Barciszewski, J. (2007). Noncoding RNAs database (ncRNAdb). *Nucleic Acids Research, 35*, D162.

Tjaden, B. (2008). Prediction of small, noncoding RNAs in bacteria using heterogeneous data. *Journal of Mathematical Biology, 56*, 183–200.

Tjaden, B., Goodwin, S. S., Opdyke, J. A., Guillier, M., Fu, D. X., Gottesman, S., et al. (2006). Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Research, 34*, 2791–2802.

Valentin-Hansen, P., Eriksen, M., & Udesen, C. (2004). The bacterial Sm-like protein Hfq: A key player in RNA transactions. *Molecular Microbiology, 51*, 1525–1533.

van Helden, J. (2003). Regulatory sequence analysis tools. *Nucleic Acids Research, 31*, 3593.

Vogel, J., Argaman, L., Wagner, E. G. H., & Altuvia, S. (2004). The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Current Biology, 14*, 2271–2276.

Vogel, J., & Sharma, C. M. (2005). How to find small non-coding RNAs in bacteria. *Biological Chemistry, 386*, 1219–1238.

Wang, C., Ding, C., Meraz, R. F., & Holbrook, S. R. (2006). PSoL: A positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics, 22*, 2590–2596.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., & Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes and Development, 15*, 1637–1651.

Yachie, N., Numata, K., Saito, R., Kanai, A., & Tomita, M. (2006). Prediction of non-coding and antisense RNA genes in *Escherichia coli* with Gapped Markov Model. *Gene, 372*, 171–181.

Zhang, Y., Sun, S., Wu, T., Wang, J., Liu, C., Chen, L., et al. (2006). Identifying Hfq-binding small RNA targets in *Escherichia coli*. *Biochemical and Biophysical Research Communication, 343*, 950–955.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G., & Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. *Molecular Microbiology, 50*, 1111–1124.