Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 276–284

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Review

Predicting sRNAs and Their Targets in Bacteria

Wuju Li [1,*], Xiaomin Ying [1], Qixuan Lu [1,2], Linxi Chen [2]

[1] *Beijing Institute of Basic Medical Sciences, Beijing 100850, China*
[2] *Institute of Pharmacy and Pharmacology, University of South China, Hengyang 421001, China*

Received 30 June 2012; revised 11 September 2012; accepted 24 September 2012
Available online 23 October 2012

Abstract

Bacterial small RNAs (sRNAs) are an emerging class of regulatory RNAs of about 40–500 nucleotides in length and, by binding to their target mRNAs or proteins, get involved in many biological processes such as sensing environmental changes and regulating gene expression. Thus, identification of bacterial sRNAs and their targets has become an important part of sRNA biology. Current strategies for discovery of sRNAs and their targets usually involve bioinformatics prediction followed by experimental validation, emphasizing a key role for bioinformatics prediction. Here, therefore, we provided an overview on prediction methods, focusing on the merits and limitations of each class of models. Finally, we will present our thinking on developing related bioinformatics models in future.

Keywords: Bacterial; sRNA; Target; Bioinformatics; Prediction

Introduction

Bacterial small RNAs (sRNAs) are an emerging class of small regulatory RNAs of about 40–500 nucleotides in length [1]. Originally they were called small non-coding RNAs [2]. However, some recent studies showed that some sRNAs, including SgrS and RNAIII [3,4], can also encode some small proteins. Thus, this class of RNA molecules is called small regulatory RNAs [5]. Through binding to their target mRNAs or proteins, these sRNAs are involved in many biological processes to regulate the expression of outer membrane proteins [6,7], iron homeostasis [8–10], quorum sensing [11,12] and bacterial virulence [13,14]. For example, RNAIII of *Staphylococcus aureus* was associated with bacterial pathogenesis [14].

The functional importance of these sRNAs in responding to environmental changes has encouraged people to find more and more sRNAs. According to the sRNA database sRNAMap [1], more than 900 sRNAs have been reported, which are mostly transcribed from the intergenic regions. sRNAs are heterogeneous in terms of sequence length and secondary structure. In addition, sRNAs are not sensitive to frame-shift or nonsense mutations. Therefore, it is still difficult to find sRNA genes directly using genetic screening methods. The current strategies often use a combination of bioinformatics prediction and experimental validation [15]. For example, through the combination of genome sequencing techniques and comparative genomics-based analysis, 88 sRNAs have been identified in the TIGR4 strain of the human pathogen *Streptococcus pneumonia* [16]. Therefore, developing prediction models for sRNA discovery is extremely critical. Up to date, two classes of prediction methods have been developed, *i.e.*, comparative genomics-based [17–22] and machine learning-based methods [23–26].

With more and more sRNAs obtained, determining their functions will also become an important part of sRNA biology. According to the locations of sRNA genes and their targets [27], sRNAs can be classified into *cis*-encoded sRNAs and *trans*-encoded sRNAs. For the *cis*-encoded sRNAs, sRNA genes overlap with their target genes and there exists a perfect base pairing region between their transcripts, while for the *trans*-encoded sRNAs, sRNA genes are separate from their target genes and there is often an imperfect base pairing region between their transcripts (**Figure 1**). For example, an imperfect base pairing region is present between

* Corresponding author.
E-mail: liwj@nic.bmi.ac.cn (Li W).

http://dx.doi.org/10.1016/j.gpb.2012.09.004

the sRNA IstR and its target mRNA tisB [5] (see sRNATarBase for detailed information, http://ccb.bmi.ac.cn/srnatarbase/). The imperfect base pairing results in much difficulty in detecting target mRNAs, which renders the experimental validation essential after computational prediction. Nevertheless, the computational methods have provided a time-saving and less labor-intensive way for the identification of sRNA targets. To this end, several prediction models have been developed [28–36].

Taken together, bioinformatics prediction plays an important role in discovering sRNAs and their targets, as pointed by some reviews on bioinformatics prediction and experimental discovery [37–40]. In the current review, we focus on the merits and limitations of each class of models and provide some perspective on future development in this field.

Prediction of bacterial sRNAs

In essence, the process of developing bioinformatics models is to learn the rules from known samples and then to apply the rules for new samples for experimental validation. Therefore, understanding the characteristics of bacterial sRNAs is vital in developing sRNA prediction models. The available literature indicates that sRNAs possess the following features [37–40]. First, sRNAs are widespread and each bacterium is assumed to contain sRNA genes. Second, sRNAs are heterogeneous in sequence length and secondary structure as mentioned previously. The sequence of sRNAs ranges from 40 to 500 nucleotides in length. Third, unlike tRNAs with the conserved cloverleaf secondary structure pattern, or eukaryotic microRNAs with similar sequence lengths and hairpin structured precursors [41], different sRNAs often have different secondary structures. Fourth, sRNAs are involved in many biological processes, such as posttranscriptional regulation of gene expression, RNA processing, mRNA stability and translation, protein degradation, plasmid replication and bacterial virulence [42–47]. The above features, on the one hand, reflect the importance of sRNAs, and on the other hand, bring difficulties in developing general models for sRNA prediction. Although many empirical models have been developed for sRNA discovery [17–26] (**Table 1**), there is little overlap between the prediction results from different models. We are still a long way from developing a perfect model for sRNA prediction.

Comparative genomics-based models for sRNA prediction

Comparative genomics-based models are a class of commonly-used methods for sRNA prediction at present. The basic assumption is that an sRNA gene should have a certain conservation of both sequence and secondary structure among a group of closely-related genomes. Therefore, how to choose the right set of closely-related genomes plays a key role in the success of comparative genomics-based models for sRNA prediction, and usually depends on the research purposes and models employed. For example, to find the sRNA genes in the intergenic regions of *Escherichia coli* [46], Argaman et al. applied the BLAST program

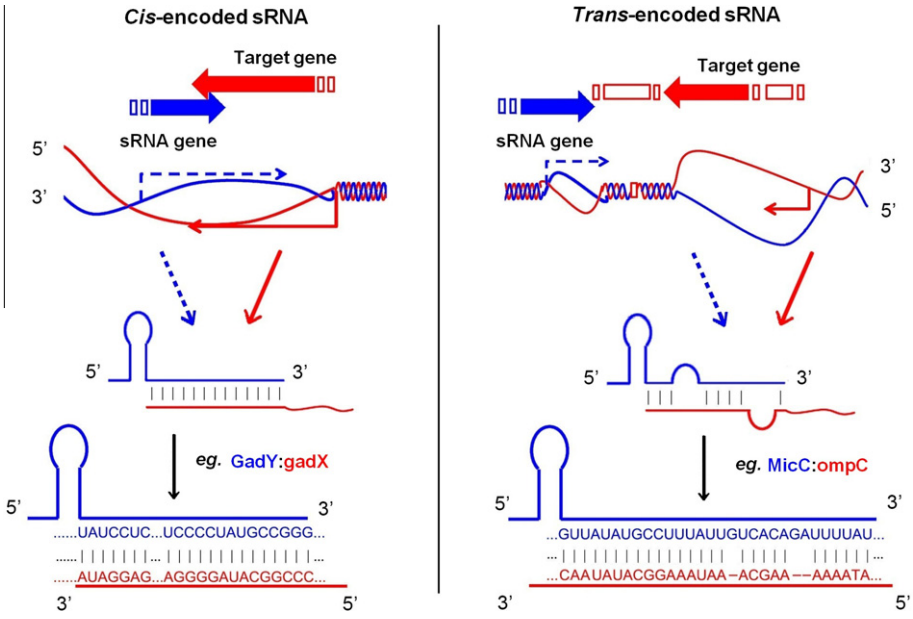

Figure 1 The action mechanisms of *cis*-encoded and *trans*-encoded sRNAs
For *cis*-encoded sRNA-target mRNA interactions, there exists a perfect base pairing region and these genes overlap but are localized on different strands. Here the interaction GadY:gadX was provided to demonstrate such interaction, in which the blue color represents sRNA and the red color stands for the target mRNA. However, for *trans*-encoded sRNA-target mRNA interactions, there exists an imperfect base pairing region. These genes are separate from each other and therefore there is no overlap between them. The interaction MicC:ompC was shown as an example. Please see sRNATarBase for detailed information (http://ccb.bmi.ac.cn/srnatarbase/). The entry names for GadY:gadX and MicC:ompC are SRNAT00067 and SRNAT00015, respectively.

to compare potential sRNA regions against the genomes of *Salmonella typhi*, *S. paratyphi* and *S. typhimurium* and identified 24 putative sRNA genes. In addition, Rivas and Eddy applied the WUBLASTN program to compare 2367 intergenic sequences of *E. coli* against the complete genome of *S. typhi* [17]. The 11,509 generated alignments were scanned using the QRNA model and finally, 33 out of 115 known ncRNAs were identified. The *E. coli* genome was also used to test the performance of the sRNAPredict program. Using sequence conservation between *E. coli* intergenic regions and *Shigella flexneri*, Livny et al. identified 50 out of 55 known sRNAs [21]. Therefore, it is very difficult to provide a general rule for how many genomes and which genomes should be included in studies of comparative genomics-based sRNA prediction.

The main steps for comparative genomics-based models to predict sRNA genes are as follows. The first step is to find closely-related genomes to a given bacterial genome. The second step is to extract intergenic regions among the selected genomes and to apply the BLAST program to compare intergenic regions pair wisely. Then, the pair wise BLAST hits are gathered into clusters of two or more sequences, and these sequence clusters are aligned using ClustalW or ncDNAlign [48]. Finally, the resulting alignments are scored using RNAz [18] or EvoFold [19]. The third step is to carry out structural conservation analysis for the intergenic regions using the above alignment. Here structural conservation means that, for some positions in each sequence, even though there is no perfect conservation of nucleotides, the base pairing information is kept. The fourth step is to predict whether the conserved intergenic regions contain the signal of promoter, transcript factor binding sites or Rho-independent terminator.

Based on some or all steps above, some programs, including QRNA [17], RNAz [18], EvoFold [19], SIPHT [20] and sRNAPredict [21], have been developed and successfully applied to finding bacterial sRNA genes. QRNA takes blast alignment of two sequences as the input, while RNAz and EvoFold take multiple sequence alignment as input, before structure analysis such as conservation and thermodynamic stability is performed to predict potential sRNA genes. Different from these tools, sRNAPredict and SIPHT only use information from blast alignment and Rho-independent terminator signal without considering structural information.

Four comparative genomics-based methods, QRNA, RNAz, sRNAPredict/SIPHT and NAPP (nucleic acid phylogenetic profiling) [22] were systematically compared using 10 sets of benchmark data in a recent evaluation paper [49], The authors found that sRNAPredict provided the best performance by comprehensively considering multiple factors such as low false positive rates, ability to identify the correct strand of sRNAs and speed of execution.

There are limitations for this class of methods. First, the aforementioned models are only applicable to the discovery of evolutionarily-conserved sRNA genes rather than the genes unique to a given genome. Second, these models are of no use if there are no closely-related genomes available for a given genome. Third, the conserved intergenic regions may contain other gene structures such as transcription factor binding sites or untranslated regions of mRNAs rather than sRNA genes. Therefore, the comparative genomics-based models are only applicable to identify some sRNA genes.

Machine learning-based models for sRNA prediction

The basic assumption of this class of models is that a given genome is composed of two parts, *i.e.*, sRNA genes and the remaining part of the genome. If we take sRNA genes as signal, the remaining part of the genome will be viewed as the background. The first step to develop machine learning-based models is to construct a training dataset including positive and negative samples. The known sRNA genes are often used as positive samples, while randomly-selected DNA sequences from the given genome are taken as negative samples. The second step is to extract features describing the samples, which is a key step in developing models. Only suitable features can improve the model performance. In addition, feature selection is also important in machine learning-based model construction. For example, in Tran's model for sRNA prediction [26], they firstly constructed a training dataset including 936 non-redundant ncRNA sequences as the positive set and the shuffled sequences of those positive samples as the negative samples. Then, they applied a t-test to find a set of features with statistical significance ($P < 0.05$) for neural network-based model construction. In fact, many feature selection methods have been applied in gene expression profile-based sample classification studies such as the Tclass system developed by our laboratory [50]. All those feature selection methods can be applied to select proper feature sets for sRNA prediction. Third, the machine learning methods such as neural networks and support vector machines are applied to develop the models. Fourth, the models developed are applied to genome-wide discovery of sRNA genes for experimental validation. If the number of predicted sRNA genes is very large, the comparative genomics-based models can be further applied to reduce the number of the genes. The main challenge in developing machine learning-based models lies in constructing training samples and features. For example, in the neural network-based model presented by Carter [23], the genetic algorithm-based model presented by Saetrom [24] and the model presented by Wang [25], the number of positive samples was enlarged by incorporating the tRNA and rRNA sequences into the training dataset.

Compared to the comparative genomics-based models, machine learning-based models for sRNA gene prediction have some advantages. For example, these models can be applied to find sRNA genes unique to a given genome. However, when we apply these models to do genome-wide discovery of sRNA genes, we often divide the genome into fragments with a certain length for prediction separately. If

the fragment is too short, it might not contain enough information for sRNA genes. Conversely, if the fragment is too large, it might contain noise information. Therefore, it is very difficult to choose the optimal window size for machine learning-based models due to the length heterogenicity of sRNA genes. Because of this, Tran et al. constructed different models using different window sizes. This might be the reason why the positive prediction value of machine learning-based models is less than that from comparative genomics-based models [26].

Prediction of sRNA targets

Developing predicting models for sRNA targets is very important. The strategy, which combines bioinformatics prediction and experimental validation for sRNA gene discovery, can also be applied to sRNA target identification [51]. To do this, understanding the features of sRNA-target interactions is the initial key step. sRNAs exert their functions through the following two ways: (1) imperfect base-pairing with their target mRNAs; and (2) binding proteins and altering their activity [43]. Imperfect base-pairing with mRNAs represents the major regulatory mechanism, which can lead to translational repression, translational activation or mRNA degradation [52]. This mechanism is the focus of current studies on sRNA-target interactions. We reviewed the related prediction models below. To date, two categories of methods, prediction models for general RNA–RNA interaction [53–65] (Table 1) and models specifically designed for sRNA-target mRNA interactions in bacteria [28–36] (Table 1), have been utilized in sRNA target discovery.

Prediction models for general RNA–RNA interactions

In essence, the sRNA-target mRNA interactions in bacteria fall into the class of RNA–RNA interactions. Therefore, the models for general RNA–RNA interaction prediction (RIP) can also be applied to investigate sRNA-target mRNA interaction.

The earliest methods for RIP are to find hybridization structure with the minimum binding free energy for two RNA molecules, using the program RNAfold [53,54] or Mfold [66] to fold the two concatenated RNA sequences. Hybridization artifacts can arise from folding the concatenation of two RNA sequences. To prevent such artifacts, many programs such as RNAcofold [54], RNAhybrid [55,56] and RNAplex [57] were presented by extending the classical RNA secondary structure prediction algorithm to two sequences. For instance, RNAhybrid [55,56] was a modification of the classic RNA secondary structure prediction method, by neglecting intra-molecular base-pairings and multi-loops. This method was originally proposed for miRNA target prediction, but it was also applied to sRNA target prediction by Sharma et al. [67]. Compared to RNAhybrid, RNAplex [57] used a slightly different energy model to reduce computational time. RNAplex performed 10–27 times faster than RNAhybrid [57].

The methods mentioned above ignore the secondary structures of two RNA molecules before they interact. To improve the prediction performance, Muckstein et al. applied a dynamic programming algorithm to search the minimum extended hybridization energy, which was defined as the sum of hybridization energy and the energy for making the binding sites accessible [68].

Since pseudo-knots were not considered in both the classical and the extensions of RNA secondary structure prediction algorithms, the aforementioned programs cannot find loop–loop interactions (kissing complex) between two RNA molecules. To address this problem, Alkan et al. presented inteRNA [59] based on joint structure of two RNA molecules. When applied in CopA-CopT and OxyS-fhlA interactions, inteRNA detected the loop-loop interactions successfully. Thereafter, multiple programs such as piRNA [60], inRNA [61], rip [62], RactIP [63], ripalign [64] and PETcofold [65] have been presented based on joint structure of two RNA molecules.

Although many programs for general RIP have been presented, most programs only provide the potential binding sites between two RNA molecules rather than determine whether two RNA sequences interact or not. In fact, two randomly selected RNA sequences can present many potential binding sites, which cannot guaranty that two RNA sequences interact. These programs are only suitable for searching binding sites given the interaction between an sRNA and a target mRNA. Therefore, it is impractical to apply these models for genome-wide prediction of sRNA targets. It is necessary to develop specific prediction models for sRNA targets.

Prediction models specifically designed for sRNA-target mRNA interactions

The first prediction model specific to sRNA-target mRNA interaction was presented by Zhang et al. [28]. They incorporated the following five features into the model: (1) Hfq-binding sites in both sRNA and target mRNA sequences; (2) flanking sequence −35 to +15 nt around the translation initiation sites in target mRNA sequences; (3) Hfq-binding sRNA structures; (4) extension alignment based on the center of loop or bulge regions from sRNA secondary structure; and (5) conservation profiles of the sRNAs and their targets among 8 closely-related organisms of *E. coli* K-12. For a given sRNA, this model scores each potential sRNA–mRNA interaction based on a modified Smith–Waterman local sequence alignment algorithm (a reward for a match and a penalty for a mismatch) and takes the mRNAs with top 10 or 50 scores as the potential targets. Among 10 experimentally-validated sRNA-target interactions, there are 7 pairs ranked in the top 50 scores. However, this model has not been applied widely because of the following reasons. First, this model was designed specifically for *E. coli* genome. For example, the conservation

Table 1 Main computational tool for prediction of bacterial sRNAs and their target mRNAs

Type	Tool	Availability	Main features	References
Comparative genomics-based models for sRNA prediction	QRNA	ftp://ftp.genetics.wustl.edu/pub/eddy/software/qrna.tar.Z	Sequence and secondary structure; suitable for two sequence alignment	[17]
	RNAz	http://www.tbi.univie.ac.at/~wash/RNAz	Sequence and secondary structure; suitable for multiple sequence alignment	[18]
	EvoFold	http://www.cbse.ucsc.edu/jsp/EvoFold	Sequence, structure and evolution; suitable for multiple sequence alignment	[19]
	SIPHT	http://bio.cs.wisc.edu/sRNA	Sequence and Rho-independent terminators	[20]
	sRNAPredict	http://www.tufts.edu/sackler/waldorlab/sRNAPredict.html	Sequence and Rho-independent terminators	[21]
	NAPP	–	Phylogenetic profiling of nucleic acid fragments; cluster analysis	[22]
Machine learning-based models for sRNA prediction	Carter et al.	http://rnagene.lbl.gov/	Nucleotide compositions and secondary structure; neural networks and support vector machines	[23]
	Sætrom et al.	–	Sequence; genetic algorithm and boosting algorithm	[24]
	PSoL	–	Sequence and secondary structure; support vector machine	[25]
	Tran et al.	http://csbl.bmb.uga.edu/publications/materials/tran/	Sequence and secondary structure; neural network	[26]
Prediction models for general RNA–RNA interactions	RNAcofold	http://www.tbi.univie.ac.at/RNA/	Extension of minimum energy folding algorithm to two sequences	[54]
	RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/ranhybrid/	Extension of minimum energy folding algorithm to two sequences; neglecting intra-molecular base-pairings and multi-loops	[55,56]
	RNAplex	http://www.tbi.univie.ac.at/~htafer/	Extension of minimum energy folding algorithm to two sequences; running faster	[57]
	RNAup	http://rna.tbi.univie.ac.at/cgi-bin/RNAup.cgi	Consideration of accessibility of binding sites	[53,58]
	inteRNA	http://www.ncrna.org/software/ractip/	Searching the joint structure of interacting RNAs with the minimum total free energy	[59]
	piRNA	http://compbio.cs.sfu.ca/taverna/pirna/	Computing the partition function over joint structures formed by two interacting nucleic acids	[60]
	Rip	http://www.combinatorics.cn/cbpc/rip.html	Computing the full partition function over joint structures formed by two interacting RNAs based on the combinatorial notion of 'tight structures'	[62]
	RactIP	http://www.ncrna.org/software/ractip/	Prediction based on joint structures using integer programming	[63]
	Ripalign	http://www.combinatorics.cn/cbpc/ripalign.html	Prediction based on joint structures with consideration of both thermodynamic stability and sequence/structure covariation	[64]
	PETcofold	http://rth.dk/resources/petcofold/submit.php	Predicting interactions and secondary structures of two multiple alignments of RNA sequences	[65]
Prediction models for sRNA-target mRNA interactions	TargetRNA	http://snowwhite.wellesley.edu/targetRNA/	Hybridization; not consider structures from sRNA or mRNA	[29,30]
	sRNATarget	http://ccb.bmi.ac.cn/srnatarget/	Sequence and RNA secondary structure profile; naïve Bayes method	[33,34]
	IntaRNA	http://www.bioinf.uni-freiburg.de/Software/	Accessibility of binding sites; user-specified seed	[32]
	RNApredator	http://rna.tbi.univie.ac.at/RNApredator	Target site accessibility; RNAup	[35]
	sTarpicker	http://ccb.bmi.ac.cn/starpicker/	Thermodynamic stability; site accessibility of sRNA and targets; naïve Bayes method	[36]

Note: The main features and properties of the related models were provided in column "Main features". For example, for QRNA, both sequence and secondary structure information were applied, and the model was suitable for two sequence alignment.

profile associated with *E. coli* was considered, which hinders people from applying the model in other organisms. Second, the model only considers secondary structures of sRNAs rather than the joint structures of two RNA sequences, which makes the model less competitive in comparison with the models presented later. Third, there is no program provided for sRNA biologists.

The second model, termed TargetRNA, was presented by Tjaden et al. [29,30]. TargetRNA included an individual base pair model and a stacked base pair model for calculating hybridization score for sRNA-target interactions. The individual base pair model was based on a modified Smith–Waterman local sequence alignment algorithm, and the stacked base pair model was a straightforward extension of RNA folding approaches with intra-molecular base-pairing prohibited, which is very similar to the statistical idea from RNAhybrid [55,56]. However, TargetRNA was optimized on a training dataset containing 12 experimentally-verified sRNA-target mRNA interactions. The optimal translational initiation region was –30 to +20 nt and seed length was 9 nt. For each potential sRNA-target mRNA interaction, the model calculates the hybridization score, which was assumed to abide by extreme value distribution. The extreme value distribution was obtained by considering a large number of randomly-generated sRNA-target mRNA interactions. Therefore, for a given sRNA, all potential sRNA-target mRNA interactions will be considered and the interactions with the top 10 or 50 smallest *P* values will be taken as the putative interactions. As a result, TargetRNA can pick up 8 from the 12 interactions with top 10 smallest *P* values.

Mandin et al. proposed a model for sRNA target prediction by searching strong sRNA-mRNA duplexes [31]. Each sRNA-mRNA duplex was scored as a sum of both positive contributions and negative contributions, which correspond to pairing nucleotides and bulges/internal loops, respectively. The cost of bulges and internal loops was empirically gauged using four validated sRNA-mRNA interactions. The statistical significance of the duplex was used as the criterion for interaction, which was assessed by comparing to an ensemble of random sequences. During prediction, the flanking regions, −140 to +90 nt around the translation initiation sites and −60 to +90 nt around the translation stop sites in target mRNA sequences, were considered.

Obviously all aforementioned models only take a certain number of top predictions (with the larger comparison scores, small free energies or small *P* values) as potential targets. To determine clearly whether a given sRNA-mRNA complex interacts or not, our group have systematically collected 46 positive samples (true interactions) and 86 negative samples (no interaction) as the training dataset. Then, according to the positions of mRNA binding sites from the validated sRNA-target mRNA interactions at that time, sub-sequences located within −30 to +30 nt of the initial start codons of targets were selected as core binding regions. Based on the hypothesis that sequences flanking the core binding regions are also likely to influence the interactions, we also extracted these flanking sequences using sliding windows. For each sub-sequence, 10 features were computed, including the percent composition of bases in interior loops, the minimum free energy (MFE) of hybridization, and the difference in the MFE values before and after hybridization. Each sRNA-target mRNA interaction was described by 10,000 features. Third, we applied the Tclass system [50] and support vector machines to construct prediction models sRNATargetNB and sRNATargetSVM, respectively [33,34]. The main difference between sRNATargetNB and sRNATargetSVM is that the former only takes six features, which were selected from 10,000 initial features using the Tclass system [50], to determine whether a given pair of sRNA and mRNA interacts or not, whereas the latter needs 10,000 features. Therefore, sRNATargetNB runs faster. Finally, the performance of the two models above was evaluated on an independent test set containing 22 positive samples and 1700 randomly-generated negative samples. Prediction accuracies are 93.03% and 80.55%, respectively.

IntaRNA was presented by Busch et al. [32], which incorporated accessibility of binding sites of two RNA molecules and a user-definable seed. Similar to RNAup [53,58], IntaRNA searched the optimal interaction with the minimum extended hybridization energy, which was defined as the sum of hybridization energy and the energy to make the binding sites accessible. The difference between IntaRNA and RNAup is that MFE values for seed regions are also included in the calculation of the minimum extended hybridization energy in IntaRNA. Three factors make IntaRNA outperform other simpler programs like RNAhybrid: (i) finding the optimal structure with the MFE; (ii) summing the energy for opening original structures of binding sites and (iii) involving the MFE of seed regions. IntaRNA provides the binding sites of two RNA molecules and the energy of the hybridization, rather than the judgment of interacting or not.

From these models, we can see that different potential binding regions are considered in different models. So, which regions are suitable for sRNA target prediction? To address this problem, we continued our efforts to collect sRNA targets in peer-reviewed papers and constructed the database sRNATarBase [5], which contains 138 sRNA–target interactions and 252 non-interaction entries. Using this database, we found that binding regions of 95.79% of the targets (91 of 95 entries containing binding regions) are located in the region −150 to 100 nt around the initial start codon of the targets. We therefore proposed another method termed sTarPicker to improve the performance of sRNA target prediction [36].

The sTarPicker method was based on a two-step model for hybridization between an sRNA and an mRNA target. The model first selects stable duplexes after screening all possible duplexes between the sRNA and the potential mRNA target. Next, hybridization between the sRNA and the target is extended to span the entire binding site. Finally, quantitative predictions are produced with an ensemble classifier generated using the Tclass system,

originally developed for gene expression profile-based sample classification by our laboratory [50]. In determining the hybridization energies of seed regions and binding regions, both thermodynamic stability and site accessibility of the sRNAs and targets were considered. The major difference between the hybridization model in sTarPicker and the one used in IntaRNA lies in the filtering of seed regions. IntaRNA does not filter any seed regions and instead, searches the optimal hybridization of two RNA molecules with the minimum extended hybridization energy in the whole length of two RNAs. sTarPicker first finds all possible seed regions, then removes the seed regions with high hybridization energy. Here we assume that only stable seed hybridization results in stable hybridization between two RNA molecules, which was verified by the real sRNA-target mRNA interactions from sRNATarBase [5].

Compared to IntaRNA, sRNATarget and TargetRNA, sTarPicker performed best in both performance of target prediction and accuracy of the predicted binding sites on 17 non-redundant validated sRNA-target pairs [36].

Recently, Eggenhofer et al. developed a webserver termed RNApredator specifically for prediction of sRNA targets [35]. RNApredator predicts sRNA targets using RNAplex [57]. To improve the prediction specificity, RNApredator also takes into account the accessibility of the target. To enable fast computation, the accessibility is pre-computed using RNAplfold [69,70]. During prediction, the web server considers the regions −200 to +200 nt of both 5′ and 3′ UTR (default) as the potential binding regions and top 100 predictions as the potential interactions.

Future thinking in developing bioinformatics models for bacterial sRNAs and their targets

Here we briefly present an overview of prediction models for bacterial sRNAs and their targets, and point out the advantage and disadvantage of each class of models. Although these models have provided much support for experimental discovery of sRNAs and their targets, they are not perfect. Here we want to emphasize three future directions in developing bioinformatics models.

The first thing is to improve the existing prediction models. Compared to methods for open reading frame identification, the prediction accuracy of sRNAs is still very low. For example, sTarPicker has the highest positive prediction value on the independent test dataset [36]; however, a large number of false positive samples were included in the prediction results. Therefore, developing better models for sRNAs and their targets is still necessary. From the perspective of statistics, we firstly need more samples. At present, some databases, such as sRNAMap [1] and Rfam [71] for sRNAs and sRNATarBase [5] for sRNA targets, have been developed. These databases provide a data source for model development. The key point is to construct suitable features to describe the bacterial sRNA gene and sRNA-target mRNA interaction. To this end, before new features are explored, it might be better to comprehensively integrate all features currently available to describe sRNAs or sRNA-target mRNA interactions. Then, different strategies for feature selection in machine-learning based model construction can be applied to search suitable features or their combinations.

The considerations mentioned above can also be applied to the second direction, *i.e.*, developing prediction models for sRNA target proteins. To our knowledge, there is no prediction model specifically for sRNA target proteins. Although the general prediction model for RNA-protein interaction can be applied here [72], we believe that models based on the sRNA-protein interaction in bacteria will provide better support for the discovery of sRNA target proteins. To this end, we have been collecting the validated sRNA-protein interactions in the database sRNATarBase [5]. However, the number of samples is so low that we are not able to develop a reliable model yet.

The third direction involves developing comprehensive bioinformatics pipelines for the discovery of sRNAs and sRNA-target interactions using high throughput sequencing technology (HTS). With the application of HTS, a large number of short reads will be generated. How to efficiently manage these short reads and to find potential sRNAs has become an important bioinformatics topic in HTS-based sRNA discovery. For example, in their recent paper [73], Pellin and his colleagues presented a bioinformatics pipeline for sRNA discovery in *Mycobacterium tuberculosis* using RNA-seq and conservation analysis, and a list of 1948 candidate sRNAs was found. Currently, HTS has been widely applied in molecular biology, resulting in the discovery of sRNA transcripts [74–81], identification of human miRNA-mRNA [82] or RNA-protein interactions [83–85] and determination of mRNA secondary structure [86–88]. However, HTS has not been applied to investigate the interactions of sRNA-protein and sRNA-mRNA in bacteria. We can predict that HTS will soon have a widespread application in sRNA biology.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

This work was supported by grants from National Key Basic Research and Development Program (Grant No. 2010CB912801), and National Natural Science Foundation of China (Grant No. 31071157 and 31271404).

References

[1] Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, et al. SRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. Nucleic Acids Res 2009;37:D150–4.

[2] Livny J, Waldor MK. Identification of small RNAs in diverse bacterial species. Curr Opin Microbiol 2007;10:96–101.

[3] Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol 2011;3.

[4] Vanderpool CK, Balasubramanian D, Lloyd CR. Dual-function RNA regulators in bacteria. Biochimie 2011;93:1943–9.
[5] Cao Y, Wu J, Liu Q, Zhao Y, Ying X, Cha L, et al. SRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. RNA 2010;16:2051–7.
[6] Guillier M, Gottesman S. Remodelling of the *Escherichia coli* outer membrane by two small regulatory RNAs. Mol Microbiol 2006;59: 231–47.
[7] Valentin-Hansen P, Johansen J, Rasmussen AA. Small RNAs controlling outer membrane porins. Curr Opin Microbiol 2007;10: 152–5.
[8] Massé E, Vanderpool CK, Gottesman S. Effect of RyhB small RNA on global iron use in *Escherichia coli*. J Bacteriol 2005;187:6962–71.
[9] Massé E, Salvail H, Desnoyers G, Arguin M. Small RNAs controlling iron metabolism. Curr Opin Microbiol 2007;10:140–5.
[10] Večerek B, Moll I, Bläsi U. Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. EMBO J 2007;26:965–75.
[11] Lenz DH, Miller MB, Zhu J, Kulkarni RV, Bassler BL. CsrA and three redundant small RNAs regulate quorum sensing in Vibrio cholerae. Mol Microbiol 2005;58:1186–202.
[12] Tu KC, Bassler BL. Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi. Genes Dev 2007;21:221–33.
[13] Romby P, Vandenesch F, Wagner EGH. The role of RNAs in the regulation of virulence-gene expression. Curr Opin Microbiol 2006;9:229–36.
[14] Toledo-Arana A, Repoila F, Cossart P. Small noncoding RNAs controlling pathogenesis. Curr Opin Microbiol 2007;10:182–8.
[15] Voss B, Georg J, Schön V, Ude S, Hess WR. Biocomputational prediction of non-coding RNAs in model cyanobacteria. BMC Genomics 2009;10:123.
[16] Acebo P, Martin-Galiano AJ, Navarro S, Zaballos Á, Amblar M. Identification of 88 regulatory small RNAs in the TIGR4 strain of the human pathogen *Streptococcus pneumoniae*. RNA 2012;18:530–46.
[17] Rivas E, Eddy S. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2001;2:8.
[18] Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 2005;102:2454–9.
[19] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2006;2:e33.
[20] Livny J, Teonadi H, Livny M, Waldor MK. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS One 2008;3:e3197.
[21] Livny J, Fogel MA, Davis BM, Waldor MK. SRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. Nucleic Acids Res 2005;33:4096–105.
[22] Marchais A, Naville M, Bohn C, Bouloc P, Gautheret D. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. Genome Res 2009;19:1084–92.
[23] Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 2001;29:3928–38.
[24] Sætrom P, Sneve R, Kristiansen KI, Snøve O, Grünfeld T, Rognes T, et al. Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. Nucleic Acids Res 2005;33:3263–70.
[25] Wang C, Ding C, Meraz RF, Holbrook SR. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics 2006;22:2590–6.
[26] Tran TT, Zhou F, Marshburn S, Stead M, Kushner SR, Xu Y. De novo computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics 2009;25:2897–905.
[27] Wagner EGH. Kill the messenger: bacterial antisense RNA promotes mRNA decay. Nat Struct Mol Biol 2009;16:804–6.
[28] Zhang Y, Sun S, Wu T, Wang J, Liu C, Chen L, et al. Identifying Hfq-binding small RNA targets in *Escherichia coli*. Biochem Biophys Res Commun 2006;343:950–5.
[29] Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, et al. Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Res 2006;34:2791–802.
[30] Tjaden B. TargetRNA: a tool for predicting targets of small RNA action in bacteria. Nucleic Acids Res 2008;36:W109–13.
[31] Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P. Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets. Nucleic Acids Res 2007;35:962–74.
[32] Busch A, Richter AS, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics 2008;24:2849–56.
[33] Zhao Y, Li H, Hou Y, Cha L, Cao Y, Wang L, et al. Construction of two mathematical models for prediction of bacterial sRNA targets. Biochem Biophys Res Commun 2008;372:346–50.
[34] Cao Y, Zhao Y, Cha L, Ying X, Wang L, Shao N, et al. sRNATarget: a web server for prediction of bacterial sRNA targets. Bioinformation 2009;3:364–6.
[35] Eggenhofer F, Tafer H, Stadler PF, Hofacker IL. RNApredator: fast accessibility-based prediction of sRNA targets. Nucleic Acids Res 2011;39:W149–54.
[36] Ying X, Cao Y, Wu J, Liu Q, Cha L, Li W. STarPicker: a method for efficient prediction of bacterial sRNA targets based on a two-step model for hybridization. PLoS One 2011;6:e22705.
[37] Vogel J, Wagner EGH. Target identification of small noncoding RNAs in bacteria. Curr Opin Microbiol 2007;10:262–70.
[38] Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. Bioinformatics 2008;24:2807–13.
[39] Backofen R, Hess WR. Computational prediction of sRNAs and their targets in bacteria. RNA Biol 2010;7:33–42.
[40] Sharma CM, Vogel J. Experimental approaches for the discovery and characterization of regulatory small RNA. Curr Opin Microbiol 2009;12:536–46.
[41] Kozomara A, Griffiths-Jones S. MiRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011;39:D152–7.
[42] Eddy SR. Computational genomics of noncoding RNA genes. Cell 2002;109:137–40.
[43] Storz G, Altuvia S, Wassarman KM. An abundance of RNA regulators. Annu Rev Biochem 2005;74:199–217.
[44] Hershberg R, Altuvia S, Margalit H. A survey of small RNA-encoding genes in *Escherichia coli*. Nucleic Acids Res 2003;31: 1813–20.
[45] Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria. Biol Chem 2005;386:1219–38.
[46] Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EGH, Margalit H, et al. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr Biol 2001;11:941–50.
[47] Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 2001;11:1369–73.
[48] Rose D, Hertel J, Reiche K, Stadler PF, Hackermüller J. NcDN-Align: plausible multiple alignments of non-protein-coding genomic sequences. Genomics 2008;92:65–74.
[49] Lu X, Goodrich-Blair H, Tjaden B. Assessing computational tools for the discovery of small RNA genes in bacteria. RNA 2011;17:1635–47.
[50] Li W, Xiong MM. Tclass: tumor classification system based on gene expression profile. Bioinformatics 2002;18:325–6.
[51] Richter AS, Schleberger C, Backofen R, Steglich C. Seed-based IntaRNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1. Bioinformatics 2010;26:1–5.
[52] Storz G, Opdyke JA, Zhang A. Controlling mRNA stability and translation with small, noncoding RNAs. Curr Opin Microbiol 2004;7:140–4.

[53] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem 1994;125:167–88.
[54] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. Algorithms Mol Biol 2011;6:26.
[55] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA 2004;10:1507–17.
[56] Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 2006;34:W451–4.
[57] Tafer H, Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. Bioinformatics 2008;24:2657–63.
[58] Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA–RNA binding. Bioinformatics 2006;22:1177–82.
[59] Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. RNA–RNA interaction prediction and antisense RNA target search. J Comput Biol 2006;13:267–82.
[60] Chitsaz H, Salari R, Sahinalp SC, Backofen R. A partition function algorithm for interacting nucleic acid strands. Bioinformatics 2009;25:i365–73.
[61] Salari R, Backofen R, Sahinalp SC. Fast prediction of RNA-RNA interaction. Algorithms Mol Biol 2010;5:5.
[62] Huang FWD, Qin J, Reidys CM, Stadler PF. Partition function and base pairing probabilities for RNA–RNA interaction prediction. Bioinformatics 2009;25:2646–54.
[63] Kato Y, Sato K, Hamada M, Watanabe Y, Asai K, Akutsu T. RactIP: fast and accurate prediction of RNA–RNA interaction using integer programming. Bioinformatics 2010;26:i460–6.
[64] Li AX, Marz M, Qin J, Reidys CM. RNA–RNA interaction prediction based on multiple sequence alignments. Bioinformatics 2011;27:456–63.
[65] Seemann SE, Richter AS, Gesell T, Backofen R, Gorodkin J. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. Bioinformatics 2011;27:211–9.
[66] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 2003;31:3406–15.
[67] Sharma CM, Darfeuille F, Plantinga TH, Vogel J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. Genes Dev 2007;21:2804–17.
[68] Muckstein U, Tafer H, Bernhart SH, Hernandez-Rosales M, Vogel J, Stadler PF, et al. Translational control by RNA–RNA interaction: improved computation of RNA–RNA binding thermodynamics. In: Elloumi M, Küng J, Linial M, Murphy RF, Schneider K, Toma C, editors. Bioinformatics research and development, vol. 13. Berlin, Heidelberg: Springer; 2008. p. 114–27.
[69] Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. J Math Biol 2008;56:129–44.
[70] Stephan B, Ullrike M, Ivo H. RNA accessibility in cubic time. Algorithms Mol Biol 2011;6:3.
[71] Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res 2009;37:D136–40.
[72] Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 2011;12:489.
[73] Pellin D, Miotto P, Ambrosi A, Cirillo DM, Di Serio C. A genome-wide identification analysis of small regulatory RNAs in mycobacterium tuberculosis by RNA-Seq and conservation analysis. PLoS One 2012;7:e32723.
[74] Oliver HF, Orsi RH, Ponnala L, Keich U, Wang W, Sun Q, et al. Deep RNA sequencing of *L. monocytogenes* reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. BMC Genomics 2009;10:641.
[75] Yoder-Himes DR, Chain PSG, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, et al. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. Proc Natl Acad Sci U S A 2009;106:3976.
[76] Camarena L, Bruno V, Euskirchen G, Poggio S, Snyder M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. PLoS Pathog 2010;6:e1000834.
[77] Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010;6:e1001090.
[78] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 2010;464:250–5.
[79] Raghavan R, Groisman EA, Ochman H. Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome Res 2011;21:1487–97.
[80] Atsuko S, Motomu M, Kiriko H, Wataru N, Reiko H, Kenji N, et al. Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. BMC Genomics 2011;12:428.
[81] Kumar R, Lawrence ML, Watt J, Cooksey AM, Burgess SC, Nanduri B. RNA-Seq based transcriptional map of bovine respiratory disease pathogen "Histophilus somni 2336". PloS One 2012;7:e29435.
[82] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 2009;460:479–86.
[83] Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, et al. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet 2008;4:e1000163.
[84] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010;141:129–41.
[85] Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol 2011;29:607–14.
[86] Mauger DM, Weeks KM. Toward global RNA structure analysis. Nat Biotechnol 2010;28:1178–9.
[87] Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods 2010;7:995–1001.
[88] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature 2010;467:103–7.