Bioinformatics and Biology Insights

REVIEW

Computational Small RNA Prediction in Bacteria

Jayavel Sridhar[1] and Paramasamy Gunasekaran[2]

[1]UGC-Networking Resource Centre in Biological Sciences, School of Biological Sciences, Madurai Kamaraj University, Madurai, TN, India. [2]UGC-NRCBS, Madurai Kamaraj University and Vice-Chancellor, Thiruvalluvar University.
Corresponding author email: jsridhar@nrcbsmku.org

Abstract: Bacterial, small RNAs were once regarded as potent regulators of gene expression and are now being considered as essential for their diversified roles. Many small RNAs are now reported to have a wide array of regulatory functions, ranging from environmental sensing to pathogenesis. Traditionally, noncoding transcripts were rarely detected by means of genetic screens. However, the availability of approximately 2200 prokaryotic genome sequences in public databases facilitates the efficient computational search of those molecules, followed by experimental validation. In principle, the following four major computational methods were applied for the prediction of sRNA locations from bacterial genome sequences: (1) comparative genomics, (2) secondary structure and thermodynamic stability, (3) 'Orphan' transcriptional signals and (4) ab initio methods regardless of sequence or structure similarity; most of these tools were applied to locate the putative genomic sRNA locations followed by experimental validation of those transcripts. Therefore, computational screening has simplified the sRNA identification process in bacteria. In this review, a plethora of small RNA prediction methods and tools that have been reported in the past decade are discussed comprehensively and assessed based on their attributes, compatibility, and their prediction accuracy.

Keywords: comparative genomics, base composition, ncRNA, sRNA prediction, structure stability, transcriptional signal

Bioinformatics and Biology Insights 2013:7 83–95

doi: 10.4137/BBI.S11213

This article is available from http://www.la-press.com.

Background

Noncoding RNA molecules (ncRNAs) are transcripts that, rather than coding for amino acids, fulfill their functions directly in the cell. These molecules are found in all life forms and regulate diverse cellular functions. ncRNAs employ a variety of mechanisms to regulate methylation of rRNA, inhibition of translation, and transcription and sequestration of regulatory proteins.[1–3] Bacterial noncoding RNAs are generally denoted as small RNAs (sRNAs). Recent advancements in bacterial sRNA research has shown that they play important regulatory roles in the expression of virulence-related factors in pathogenic bacteria,[4,5] as well as in controlling pathogenicity,[6] mediating Iron-Response associated virulence,[7] and host-induced expression in virulence of *Salmonella typhimurium*.[8] Prokaryotic sRNAs were also found to be involved in regulating the Cell to Cell Communication in quorum sensing of *Vibrio harveyi*[9] and photo oxidative stress response in *Rhodobacter sphaeroides.*[10] Available literature articulates the numerous genomic screens for 'novel' sRNAs and functional characterization studies in enterobacterial model organisms such as *Escherichia coli* and *Salmonella typhimurium,*[11] with varying level of success. One third of the known sRNAs in *Salmonella typhimurium* have been functionally characterized to regulate the outer membrane protein and membrane transporters which emerge as a functional network.[11] Recent studies have also established the indispensable nature of sRNAs in cell adaptation, survival, and pathogenesis.[4] The number of noncoding sRNAs are growing and are being assigned with many unexpected functional roles in bacteria; this has resulted in an urgent need for efficient computational platforms for their annotations in genome projects. There is therefore considerably anxiety in the search for sRNAs in bacteria. The advent of genome sequencing data has supported many computational investigations for sRNAs followed by in vivo validations in bacteria,[12–20] Most of the known sRNAs reported in literature and Rfam databases[21] were tracked through bio-computational genomic screens in model organisms such as *Escherichia coli* and *Salmonella typhimurium*. Traditionally coding genes (tRNAs, rRNAs) were mainly annotated using automated pipelines[22] and noncoding sRNA regions were overlooked for the past fifty years.[23] The number of completely sequenced bacterial genomes is increasing and the corresponding functional annotation is becoming increasingly more difficult for biologists. Generally, the genes coding for proteins are identified through diverse algorithms designed to identify a set of transcription factor binding sites and signals in the DNA sequences. However, the use of computational utilities to discover bacterial sRNAs is not as easy a task as originally expected, primarily because (A) sRNA regions are diverse in length, ranging from 50–500 nts; (B) there are no common secondary structure, such as the clover leaf model of tRNA; (C) sRNAs do not exhibit any statistically distinguishable nucleotide biases; and (D) a lack of sRNA conservation among distantly related genomes. Many powerful attempts were made to crack the code for predicting the genomic locations of the sRNAs in bacteria; these attempts involved implementing and searching the sRNA specific properties collected from literature. The RNA sequence homology, thermodynamically favorable secondary structure (using free energy models), structure similarity searches, consensus secondary structures, and comparative genomics, are all frequently implemented, either alone or in combination with the above methods, in many sRNA finding tools designed to locate the appropriate sRNA regions. Additionally, 'Orphan' promoters, terminators, di/tetra-nucleotide frequencies, and a high level of secondary structure conservation were regarded as unique features of sRNAs and implemented in the design of computational sRNA identification tools.

Comparative genomics-based sRNA identification involves the comparison of sRNA specific inputs against the entire genome sequence data, using definitive protocols for similar sequence or structures.[24] The secondary structure of sRNA, virtually represented in the form of a RNA descriptor, can be scanned against the entire genome sequence using RNAMotif,[25] inclusive of RNAMOT and RNABOB,[26] or specific scripts written in *palingol*. The RNA structure similarity and comparative genomics are routinely employed to identify sRNAs in most sequenced bacterial genomes. The comparative analysis of RNA secondary structures can also be performed using co-variance based methods encompassed in tools such as QRNA.[14] Furthermore, these secondary structure comparison methods can be compiled into two categories, namely thermodynamic

stability involving RNALfoldz,[27] and structure consensus involving tools such as RNAZ.[16] Among this group, RNAZ employs both thermodynamic stability and structure consensus to predict probable sRNA regions. Transcriptional signal-based sRNA prediction tools include sRNApredict,[18] sRNAscanner,[19] and sRNAfinder,[29] which uses either genomic DNA sequence for signal prediction, or pre-computed signal coordinates from other databases. Generic methods such as nocoRNAc and smyRNA make use of antisense RNA transcript expression-profiling and ab initio structural sequence motif-based discovery, respectively.[20,30] Such collective efforts have lead to an explosion in the number of predicted sRNAs in diverse bacterial genomes; while these are documented in specialized databases such as Rfam,[21] most still require experimental validation. The experimental protocols used in biochemical sRNA validations[31] and functional characterizations[32] have been extensively discussed by earlier reviews. The main objective of this review is to update the available computational approaches for prokaryotic sRNA predictions with regard to their functionality, developmental background, sensitivity, specificity, success rate, and their ease of use. Glimpses of the software tools developed for the identification of sRNAs are also analyzed for their advantages and disadvantages, their current status, and their future prospects.

Computational Small RNA Prediction

Generally, the primary structure or sequence of RNA can be represented in the form of a series of nucleotides (A, U, G and C). The RNA secondary structures are composed of stems, loops, hairpins, and bulges; these are virtually described through distinct RNA descriptors. Both the primary RNA sequence and virtual RNA descriptors are used as inputs in efforts to identify similar sRNA regions in partial or complete genome sequences. Alternatively, the unique properties and features of sRNAs can be used in the computational screening of sRNA regions by evaluating the sequences at large scale. Most of the recently reported computational sRNA prediction methods employ (1) the principles of comparative genomics, (2) thermodynamically favorable secondary structures, (3) transcriptional signals, and (4) ab initio methods using the sRNA specific features. Computational protocols employed in the four major sRNA prediction approaches are shown in Figure 1.

Comparative Genomics Based Tools

In the past decade, comparative genomics has been extensively used for the identification of sRNAs in several bacterial genomes such as *Streptomyces,*[33] *Cyanobacteria,*[34] *Sinorhizobium meliloti,*[35] *Francisella tularensis,*[36] *Xanthomonas oryzae* pathovar *oryzae,*[37] and *Clostridium* sp.[38] Many computational tools were developed with the comparative genomic principle in order to screen sRNAs from genomic sequences. Among them, QRNA[14] is one of the significant attempts at employing consensus structure analysis in combination with comparative genomics to identify sRNA regions in bacterial genomes. In principle QRNA works with three probabilistic models to detect CODING regions, RNA regions, and the NULL hypothesis model. Furthermore, QRNA applies covariance-based mutation analysis for predicting noncoding RNA regions. It acts as a prototype for most of the presently available ncRNA prediction tools. The methodology has adopted the detection of conserved structural RNAs and *cis* regulatory RNAs. The SCFG (Stochastic Context Free Grammar) algorithm is composed of three pair-HMMs (Hidden Markov Models)—namely RNA, COD, and OTH—for determining mutations in aligned sequences. OTH depicts NULL hypothesis of RNA as being either coding or noncoding. The COD model corresponds to substitution mutations for confirmation of coding regions and the RNA model pertains to RNA secondary structure conserved by a mutation pattern. The sRNA or RNA probabilistic model was designed to identify covariances in stem-loop structures by implementing with the above SCFG model. Scoring is calculated based on Bayesian posterior probability, which leads to the identification of candidate ncRNA genes.[14] The QRNA approach gives more weight to the intergenic conservation among related genomes as an indicator of a probable sRNA region. It has led to the development of tools that detect consensus RNA structures and motifs from multiple sequence alignments. ERPIN (Easy RNA Profile IdentificatioN) is an algorithm used to define RNA motifs by using multiple sequence alignments and secondary structure consensus. A log-odds score profile of each helix and single strand in the multiple sequence alignment

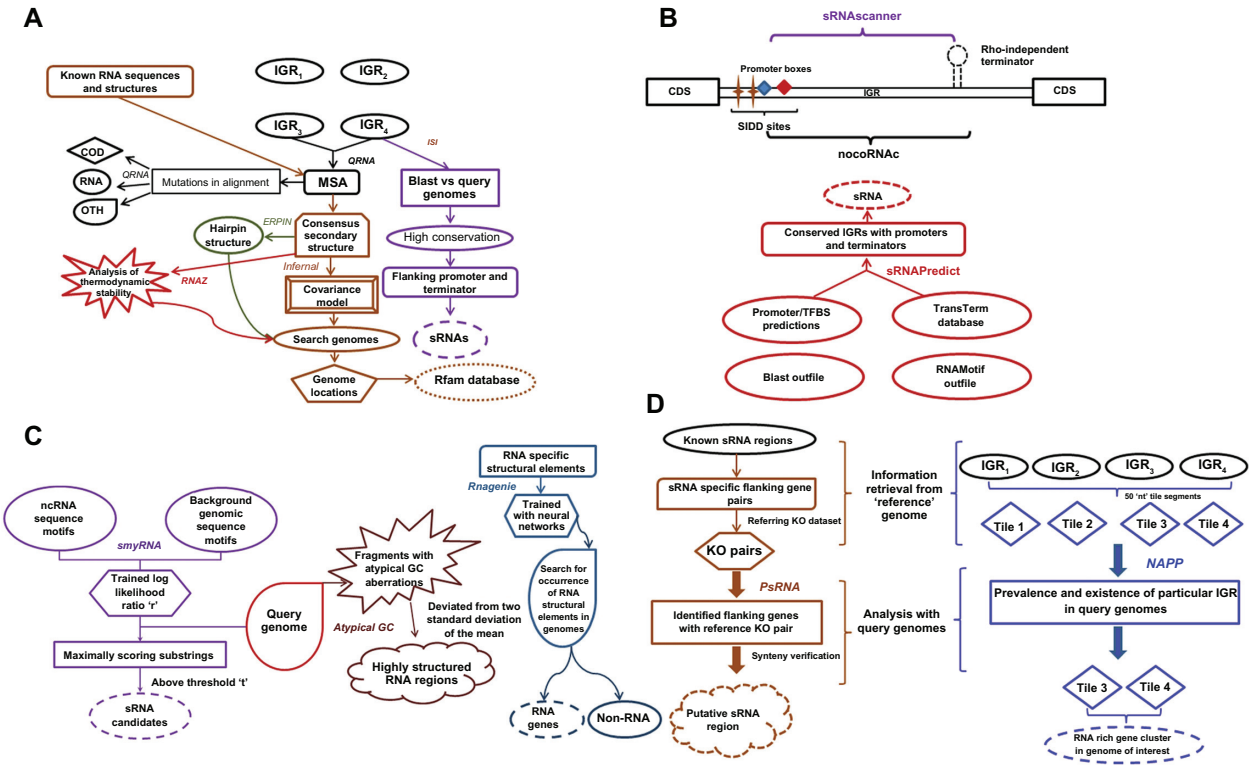

Figure 1 (**A**) Comparative genomics based protocols utilized in the computational sRNA prediction tools: *QRNA, ERPIN, ISI* and *RNAZ*; (**B**) methodology adapted in the transcriptional signal-based sRNA finders: *sRNAscanner* and *sRNAPredict*; (**C**) sequence based ab initio sRNA detection methods: *Atypical GC, RNAGENiE* and *smyRNA*; (**D**) non-sequence based ab initio sRNA detection methods: *PsRNA* and *NAPP*.
Abbreviations: IGR, InterGenic Region; sRNA, small RNA; rRNA, ribosomal RNA; tRNA, transfer RNA; CDS, Coding Domain Sequence; KO, KEGG Orthology; TFBS, Transcription Factor Binding Site.

defines the purpose of prediction. A dynamic programming square matrix detects the presence of hairpin structures among the consensus secondary structure. The ERPIN server computes E-values for every log odd score profiles as that of BLAST. ERPIN is freely available online (http://tagc.univ-mrs.fr/erpin).[39] MSARI is a program developed for the identification of noncoding RNA by detecting the RNA specific consensus secondary structure among the multiple sequence alignments. MSARI utilizes RNAFOLD[40] to generate RNA secondary structure in multiple sequence alignments and CLUSTALW[41] to make sequence alignments. A reverse complementary approach based on a Bonferroni-style test is involved in elucidating null hypothesis distributions. MSARI can be used in comparative search of ncRNA orthologs, among the related organisms with conserved RNA secondary structure, and statistical estimation of mutations in multiple alignments.[42] Both the ERPIN and MSARI utilities accept multiple sequence alignments as inputs to detect probable consensus structures of RNA from the sequence alignment files. To facilitate the global search of reliable RNA consensus structures from genome sequence databases, a computational tool known as INFERNAL (INFERence of RNA Alignment) was developed which scores combination of both sequence and structure consensus.[43,44] INFERNAL works on an HMM based covariance model for building secondary structure consensus from a RNA family and searching sequence databases for RNA structure and similarities. INFERNAL was implemented with special SCFGs in the form of Covariance Models (CMs). CMs are used to search against the genome sequences for consensus structure/sequences of particular RNA groups and to assign scores. Generally, INFERNAL performs a homology search for putative ones, followed by secondary structure based multiple sequence alignments. The latest release of INFERNAL 1.0.2 is inclusive of four programs, namely *cmbuild*, *cmcalibrate*, *cmsearch* and *cmalign*. INFERNAL is more capable of detecting conserved secondary structures, though they do not have any sequence similarity with the training CMs. An E-value is assigned for

RNA alignments and thus predicting the locations of the noncoding RNA.[44] INFERNAL is used mainly in the development of Rfam, a database of RNA alignments and CMs.[21,43] The CMs generated by the Rfam database are searched against the genome databases, primarily in conjunction with the INFERNAL tool, for homologs of known structural RNA families. The detection of numerous sRNAs (almost ~90%) in the intergenic regions created awareness among the researchers of design tools that look for sRNAs, in particular in the intergenic regions. Subsequently, a tool known as Intergenic Sequence Inspector (ISI) was developed to predict regulatory sRNAs by analyzing the intergenic conservation of the sequences among the phylogenetically related species, displaying RNA structural features flanked with putative promoters and terminators.[17] ISI is a PERL package which requires 'Bioperl 0.9.3' modules and 'NCBI Blast' utilities to perform the searches. The IGR extractor utility of ISI was used to extract the 'empty' intergenic regions from the reference genome based on their annotations and blasted against the available bacterial genomes. The blast results can be analyzed and sorted by 'Blast analyzer' according to their 'Expected value'. Furthermore, the high scoring individual alignments of the bacterial IGRs can be converted into multiple alignments to look for their conservation level. Multiple sequence alignments showing high level of sequence conservation and flanked by upstream promoters and downstream terminators are predicted as putative sRNAs. Finally, ISI selects and visualizes candidate IGRs bearing conservations and RNA signatures, along with transcriptional signals located in the analyzed genome. ISI has retained most of the known sRNAs and predicted new sRNAs in the *E. coli* K12-MG1655 genome. ISI is available for download online from http://www.biochpharma.univ-rennes1.fr/.

Secondary Structure Based Methods

A comparatively efficient and fast way to get information on a RNA molecule is its secondary structure. Reliable secondary structure prediction is a prerequisite for most types of computational RNA analysis. Consensus structure prediction among a set of RNA molecules is a routinely used process to infer their RNA families and is performed for tRNA and tmRNA. Likewise, consensus structure prediction is also an ideal starting point for sRNA prediction; however, it has to be validated by a suitable measure of significance like thermodynamic stability. RNAMotif is a tool developed to search a database for a motif which pertains to a particular secondary structure interface. The RNA motif of interest is provided in the form of a RNA descriptor file; the RNAMotif algorithm will scan the entire genome database and precisely locate the specified motif with their coordinates. RNAMotif uses the 'novel' RNA descriptor as an input to define the secondary structures of RNA which are inclusive of hairpins, properly nested hairpins, pseudoknots, and other structural elements which yield a score based on the coding nature of the nucleotide. RNAMotif algorithm consists of two stages namely, (1) construction of RNA structure descriptors and (2) searching/scoring.[25] Execution of the search in databases for pattern matches and scoring of the predicted RNA regions are based on the nearest-neighbor energy system proposed by Mathews et al.[45] These computational approaches were either applicable or limited only by the known secondary structure of a specific sRNA family.

Consensus RNA Structures and Thermodynamic Stability Based Approaches

Most of the existing methods based on RNA structure analysis have shown minimum reliability when the evolutionary distance between the two sequences lies outside of the optimal range. The secondary structure of sRNAs predicted through computational methods does not depict the actual functionally active structures and such secondary structure prediction alone is not sufficient for sRNA screens.[14] To get functionally active sRNA structures, the sRNA should be of a thermodynamically favorable minimum free energy (MFE) secondary structure which has lower free energy than random sRNA sequences. To achieve maximum enhanced accuracy, large numbers of multiple sequence alignments are essential to detect reliable RNA hits. However, constraints in getting a large number of datasets limits the use of RNA-structure-based methods in the genomic screening of sRNAs. To overcome these limitations, an additional measure MFE was added to the structure conservation in sRNA screens and implemented in a series of computational tools.[16,27] The main objective of evaluating the thermodynamic stability of RNA structures

is to efficiently detect functional sRNAs in multiple sequence alignments.

One of the more prominent tools, RNAZ is designed specifically for the efficient detection of sRNAs in alignments generated from only a few input sequences of genomes with few related sequences. It is a PERL based software which utilizes both secondary structure consensus and thermodynamic stability as measures of predicting noncoding RNAs with high specificity and sensitivity. Initial fold of the single sequence is computed using RNAFOLD and consensus folding of the aligned structures is performed with RNAalifold[46] using the same parameters. Washietl et al[16] have applied a novel SVM-based regression analysis with synthetic sequences of different length and composition in order to optimize the z-score calculation. The method initially employed MFE of RNA folding and z-scores of regression followed by SVM-based classification to differentiate whether the RNA belongs to the sRNA or not. The approach is also a part of the comparative genomics-based analysis of annotation reported in RNA databases such as Rfam. Presently, RNAZ is available (http://www.tbi.univie.ac.at/wash/RNA) for prediction of ncRNAs and *cis* regulatory RNAs. The method is applicable for large scale genomic screens and aligned sequences as well.[16]

Another similar tool developed was RNALfoldz, an expanded version of the RNALfold algorithm which can predict sRNAs based on local secondary structure, along with their thermodynamic stability. Detection of functional sRNA structures is carried out by the RNALFOLD algorithm. The MFE of the functional sRNA structures is compared to the MFE of the shuffled sequences having equal length and %GC composition. RNALfoldz can evaluate the thermodynamic stability of the predicted secondary structures by estimating the z-score and using the Support Vector Regression (SVR) introduced by Washietl et al;[16] the RNALfoldz algorithm also has local RNA secondary structure prediction and the ability to efficiently search for thermodynamically stable sRNA structures. The z-score computed by this approach is depending on the threshold for test sequence and the random true positive and false negative sequences.[27] RNALfoldz has demonstrated its applicability in detection of the thermodynamically stable functional of sRNAs in its application in the genome sequences of *E. coli*.

Yet another similar tool is CARNAC, which can be used for predicting families of homologous noncoding RNAs based on secondary structure information. It uses three distinct parameters, namely energy minimization, phylogenetic comparison, and sequence conservation. The CARNAC algorithm is composed of three stages: (1) identification of potential stems; (2) analysis of all sequences to build a pair-wise folding; and (3) preparation of a stem graph. The set of single stranded RNA sequences that need not to be aligned is accepted as input; the folding relies on the thermodynamic model with energy minimization. The CARNAC web server is written in ANSI C and freely accessible at online (http://bioinfo.lifl.fr/carnac).[47]

Transcriptional Signal-Based Methods

Most of the transcripts are expected to be encoded by the free standing genes in intergenic regions and encompassed by transcription factor binding sites and/or promoters and terminator signals. However, tracing the occurrence of transcriptional signals in intergenic regions is a difficult task. Moreover the transcriptional signals are rated as weak compared to the signals of the coding genes. It is presumed that most of the sRNAs expressed under stress conditions might be controlled by the rare transcriptional promoters. However, few sRNA identification tools have been developed by utilizing the 'Orphan' transcriptional signals in the intergenic regions, in order to predict sRNAs. The available promoters and rho-independent terminators of the coding genes were first used by Chen et al[15] to detect the probable sRNA regions in the intergenic regions of *E. coli*. This study looked for the σ^{70} promoter within a short distance of a rho-independent terminator in the intergenic regions; it predicted 144 non-translatable sRNAs in *E. coli*, with few of them experimentally validated through northern analysis. Unfortunately, the computational scripts used in this study were not provided to users. Since original development of these tools, a few tools have been developed using different statistical models. The sRNAPredict is a first co-ordinate based algorithm designed to predict the putative sRNA regions in bacteria[18] by using the locations of transcriptional signals. sRNAPredict depends on the promoter signals, transcription factor binding sites, rho-independent terminator signals predicted by TRANSTERMHP,[48] and BLAST[49] outputs

as predictive features of sRNAs. It uses coordinate based algorithms to integrate the respective positions of individual predictive features and to locate the sRNAs in the intergenic regions. The entire program was built in C++ and aimed to locate intergenic and 5′ or 3′ sRNAs. The program uses the predictive features obtained from other databases (eg, TRANSTERMHP) or output files of RNAMotif for the prediction of rho-independent terminators, TRANSFAC data for transcription factors, and BLASTN 2.0 outputs for the selection of conserved IGR coordinates used to predict probable sRNA regions; sRNApredict2 and sRNApredict3/SIPHT are recent versions of the sRNApredict suite that are used in the efficient prediction of sRNAs, with a high level of specificity. SIPHT is a compatible web version of sRNAPredict3 that searches approximately 1900 bacterial replicons from the NCBI database and predicts putative sRNA locations. sRNApredict3 is inclusive of sequence comparable options to look for conserved sRNAs, along with an earlier coordinate based approach.[50] Transcripts of the six out of the nine selected sRNA candidates were detected through northern analysis and confirmed their expression. Instead of simply looking for the intergenic signals, Tjaden's group has developed a tool known as sRNAFinder to detect sRNAs by combining the high-throughput experimental data with their relative transcriptional signals as predictive features of sRNA identification. sRNAFinder uses the multi-probabilistic method to identify the noncoding sRNAs in prokaryotic genomes. It has been implemented with nine state (four on the positive strand, four on negative strand and an intergenic state) transitions in the Genaralized Markov Model (GMM) system to predict the sRNAs. The GMM incorporates heterogeneous data, including primary sequence, transcript expression data from microarrays, conserved RNA secondary structures identified from comparative studies along with promoter, and terminator information. sRNAFinder also makes use of the SCFG (Stochastic Context Free Grammar) model used by QRNA and transcript expression profile from microarray data; it predicts the probability in confidence of interval based on existing sequence annotations. Like sRNApredict, sRNAFinder uses comparative genomics information for the purpose of predicting noncoding RNA genes.[29] The above transcriptional signal-based tools depend on the promoters, transcription factor binding sites, and the terminators predicted or available from other databases. Unfortunately, predictive features were not available for all the available genome sequences in these databases, which therefore restricts sRNA predictions using these tools. To avoid those limitations, a generic platform sRNAscanner[19] was developed to predict sRNAs and which computes the locations of the intergenic signals using the given family specific training data sets. sRNAscanner is a generic transcriptional signal-based computational method using a Positional Weight Matrix (PWM)-based strategy for the discovery of intergenic sRNA transcriptional units (TUs) in completely sequenced bacterial genomes. The main advantage with sRNAscanner is that it computes the predictive features on its own; it uses its own algorithm and the training PWM dataset to calculate the genomic locations of the promoter, transcription factor, and terminator signals. Unlike the sRNApredict2 series of suites which depend on the predictive features retrieved from other databases, sRNAscanner opens the possibility of user specific training PWM construction and sRNA identification. The sRNAscanner consists of algorithms to perform the following functions: (a) construct PWMs from sRNA-specific transcriptional signals; (b) search complete genome sequences using constructed PWMs that identify intergenic promoter/transcription factor binding sites and terminator locations; (c) perform coordinate based integration of promoter/terminator signals to define putative intergenic transcriptional units (TU); and (d) select predicted TUs based on cumulative sum of scores (CSS) values above a user defined threshold.[19] Moreover, the sensitivity and specificity profile of sRNAscanner was first evaluated through the Receiver Operator Characteristic (ROC) curves and confirmed its satisfactory performance. Six out of the sixteen sRNA candidates have yielded distinct northern-detectable transcripts of similar sizes, as per sRNAscanner predictions. Furthermore, the 5′ ends of the above six transcripts were also marked through 5′RACE experiments. Thus, sRNAscanner was proved as an efficient platform for the prediction of sRNAs in any bacterial genome, provided it was provided with family specific training sets.[19]

To advance the functional characterization of the identified sRNAs, specific tools were developed to

find the interactions between sRNAs and mRNAs. Recently, an attempt was made with the nocoRNAc tool to study the sRNA-mRNA interactions together with with the sRNA predictions from the genomes. The nocoRNAc is a Java based program for the genome wide prediction and characterization of ncRNA transcripts. nocoRNAc incorporates a set of protocols for the detection of transcriptional features which are then integrated to determine the sRNA transcript coordinates. The nocoRNAc program uses the transcription termination signals from TRANSTERMHP; promoters are identified using the SIDD model (Stress Induced Duplex Destabilization). Subsequently the program searches for the known RNA motifs from the Rfam database. IntaRNA tool was also implemented in the nocoRNAc package to predict sRNA-mRNA interactions so as to elucidate the regulatory role of the sRNAs predicted by nocoRNAc.[20]

Ab initio sRNA finders

Existing biochemical and computational studies have reported many predictive sRNA features. They are applied for the identification of sRNAs based on preferential occurrence of RNA specific structural elements, di/tri nucleotide preferences, and atypical GC properties of the genome sequence information. In the early 2000s sRNA specific features were identified by Carter et al[51] and applied in the search for similar sRNAs in *E. coli*. This method was implemented in the RNAGENiE tool,[51] which locates new RNA regions based on the finding that most of the functional RNAs (fRNAs) share common secondary structural elements like double helices, uridine turns, UNCG tetraloops, GNRA tetra loops, tetraloop receptors, adenosine platforms, and non-Watson Crick mis-pairs in a symmetric internal loops. Known RNA-specific structural elements were trained with neural networks to recognize RNA genes in *E. coli*. RNAGENiE looks for the preferential occurrence of the above secondary structural elements in the query genome sequences and differentiate them into RNA and non-RNA genes. Additionally, sequence based descriptors were also used to differentiate RNA genes from non-RNA genes. RNAGENiE achieved a greater accuracy by using a second set of parameters consisting of known RNA sequence motifs and the calculated free energy of folding.[51] Thus, the combination of RNA secondary structure prediction with base composition statistics trained with neural networks has been proposed to predict functional RNAs; however, their reliability is questionable due to lack of experimental validations. The RNAGENiE interface is available for users and opens the possibility of checking the query sequence as RNA or non-RNA genes (http://rnagenie.lbl.gov/).

After the initial studies, the RNA regions and non-RNA regions were surveyed for their sequences motifs and applied for sRNA screens. SmyRNA is an ab initio ncRNA gene finder that utilizes differential distributions of sequence motifs between ncRNAs and background genome sequences. Based on the trained log-likelihood ratio 'r', smyRNA can locate other ncRNAs on an input genome sequence 'G' by determining the maximally scoring substrings of the input sequence 'G'. Those substrings whose score is over a user defined threshold 't' are then declared as ncRNA candidates. Using a k-mer motif, log-likelihood scores for a specific sequence to be in a potential ncRNA sequence is computed. Finally, the maximum scoring subsequences of a genome, which can then be considered as a candidate ncRNA gene, is identified.[30] In general, the RNA regions are shown to have high %GC compared to the background genomic %GC. It is established and applicable in the search of sRNAs in the thermophiles. The highly structured RNA regions are expected to have high GC% variations with CDS or 'empty' intergenic regions.

The AtypicalGC tool identifies the regions showing atypical GC aberrations out of the 40% to 60% interval. It considers the entire genome as multiple fragments dependant on the user given window length and computes GC% on a specified length of genetic fragment, from the center of the sliding window pointing position. The regions showing more deviation, measured from two standard deviations of the mean, are considered atypical regions.[52] Most of the sRNAs identified through sequence homology and structure consensus methods are located in the specific intergenic regions showing phylogenetic conservation. Even the non-homologous sRNA regions identified in earlier studies were also found in those phylogenetically conserved regions.[53] Based on these intergenic conservation profiles, a computational protocol was developed to compute sRNA elements in bacteria. Both NAPP and PsRNA methods

efficiently predict the intergenic regions of sRNAs without any precise start and ends. Due to this unique nature, we have categorized them under the ab initio group, though they apply adjacent flanking gene or gene cluster information of known sRNAs. NAPP (Nucleic Acids Phylogenetic Profiling) is a clustering method that efficiently predicts the noncoding sRNA elements in bacterial genomes. NAPP works on two computational aspects: (1) information retrieval from reference genome; and (2) identification of putative sRNA regions in query genomes. It converts the intergenic regions of the reference genome into 50 nt 'tile' segments and then compares the prevalence and existence of the particular intergenic region in the available genome sequences. If the particular intergenic 'tile' segment is found it will be classified according to their evolutionary distance. Traditionally a group of sRNA 'tiles' always clusters together with certain types of CDSs, which yields important clues on the functions associated with these sRNAs. NAPP is preferable for the users to retrieve RNA rich gene clusters from the genome of interest. NAPP is available online (http://rna.igmors.u-psud.fr/NAPP/index.php).[54] The earlier phylogenetic profiles have confirmed the precise positioning of sRNAs with their specific flanking genes.[53] The available controlled vocabulary to define the functional ortholog groups of genes, eg, the KEGG Orthology (KO) numbers, support the search of particular phylogenetic intergenic profiles using their conserved flanking gene pairs. The availability of specific flanking gene pairs is utilized by PsRNA to identify the putative sRNA regions of the known groups. PsRNA is a computing engine used to identify putative sRNA locations within the intergenic regions of the bacterial genome of interest. PsRNA is an automation of the earlier sRNA identification strategy which uses conserved flanking gene synteny and genomic backbone retention information.[53] It uses the functional assignment of the sRNA specific conserved flanking genes in order to identify similar RNA regions in the query genomes, even in the absence of sequence homology. PsRNA has used the KO numbers as controlled vocabularies to identify the putative sRNA locations in the query genomes. PsRNA scripting consists of two parts: (1) information retrieval from the reference genome; and (2) the search for sRNA locations in the query genomes. The user given sRNA information (Id, coordinate, or flanking gene pair ids) will be analyzed by the PsRNA server and their corresponding flanking gene pairs will be converted into their corresponding KO id pairs (if found). Furthermore, the KO pair will be loo at the query genomes for their coexistence and synteny retention. If any of the particular flanking gene pairs is identified with similar KO pairs and satisfies the synteny criteria, it will be proposed as putative sRNA region. PsRNA server was tested with the 22 enterobacterial genomes and is currently available online (http://bioserver1.physics.iisc.ernet.in/psrna/).[55] The PsRNA method is applicable solely for the comparative analysis of a known sRNA group among related genomes. It cannot be used for the identification of 'novel' sRNA groups. A summary of the various computational tools used for sRNA prediction in bacteria is available in Table 1.

Future Perspectives

In the last few years, many computational approaches have been developed to detect ubiquitous bacterial noncoding sRNAs; a few approaches such as base composition statistics, sequence, and structure conservations, have already been reviewed.[56] Among them, comparative genomics has predicted and marked the existence of a plethora of sRNAs, and many of their expressions have been confirmed in vivo. A few of the conserved backbone or intergenic 'tile'-based comparative approaches have predicted the non-homologous sRNA regions, even without any sequence homology; this clearly shows the significance of these approaches in tracing the sRNA regions. Though many sRNAs have been identified through comparative approach-based tools, transcriptional signal-based tools are very much promising in detecting 'novel' intergenic sRNAs. Transcriptional signal-based methods have shown their prowess in predicting sRNAs, even in gram-positive bacteria. In most of the cases the sigma^{70} promoter signals are routinely used to identify the 'Orphan' promoter signals in the intergenic regions; the inclusion of other stress responsive rare promoters, ie, sigma,[54] sigma^{27} and sigma,[38] in the positive training set could predict additional stress responsive sRNAs in the future. Such transcriptional signal-based methods work perfectly only with the detection of intergenic sRNAs and they are difficult to apply in the detection of untranslated

Table 1. Summary of the various computational methods applied for sRNA prediction in bacteria.

S. No.	Method	Tool	Properties
1	Comparative genomics	*QRNA*	It applies SCFG to test and differentiate the alignments in to: COD, RNA and OTH models
		ERPIN	Reads multiple sequence alignments and secondary structures to infer Secondary structure profile (SSP)
		ISI	Search sRNAs based on intergenic conservation (IGR), RNA structural features and terminators
		INFERNAL	HMM based covariance model (CM) was used to build RNA secondary structure and search
		MSARI	Detects RNA specific common stems from multiple sequence alignments using distribution-mixture method
	RNA structure and thermodynamic stability based methods	*RNAZ*	RNAZ applies SVM based structural regression analysis to compute z-score and differentiate the minimal free energy structures
2	Transcriptional signal based sRNA finders	*sRNAscanner*	Generic sRNA finder applied for any genome with specific training data
		*sRNAPredict*3/ *SIPHT*	Coordinate based algorithms to integrate the locations of promoters/TFBS, terminators along with sequence conservation
3	Sequence dependent Ab-initio sRNA detection methods	*Atypical GC*	Compute G and C content of a particular position using sliding window and predicts RNA regions
		RNAGENiE	Known RNA structural elements were trained with neural networks and applied to differentiate RNA and non-RNA genes
		smyRNA	Utilizes differential distributions of sequence motifs between ncRNAs and background genome sequences
4	Sequence independent Ab-initio sRNA detection methods	*PsRNA*	It uses KEGG orthology numbers of the flanking genes to locate the sRNA specific intergenic regions
		NAPP	IGR's of reference genome are tiled into 50 nt segments and classified based on their occurrence profile in 1000 genomes

region (UTR) encoded sRNAs or riboswitches. Most of the structure consensus-based methods utilize thermodynamic stability and conservation of the predicted transcripts to identify the locations of putative sRNAs. Though the preferential occurrence of specific sequence motifs and %GC of ncRNA regions are used for the prediction of sRNAs, it is nevertheless difficult to identify a commonality or particular statistical bias among the prokaryotic sRNA groups.

Most of the sRNA prediction algorithms/tools were evaluated for their capacity in retaining the known sRNAs in a given reference organism.[19,50] The prediction accuracy of these computational

Advantages	Disadvantages	Availability/website	Reference
First systematic method for ncRNA detection among closely related organisms. Intergenic conservation is considered as indicator of sRNA regions	Restricted to pairwise alignments alone	http://selab.janelia.org/software.html	14
Complex RNA descriptors are not required. Dynamic programming was applied to search helix and hairpin structures (SSP) with log-odd score and E-value	Multiple sequence alignment and consensus structures are mandatory	http://tagc.univ-mrs.fr/erpin/	39
ISI has retained many sRNAs in *E. Coli.* Usage with perl and Bioperl modules	Conserved IGRs without flanking promoters and terminators are missed	http://www.biochpharma.univ-rennes1.fr/	17
CM based search of particular RNA against genomes are computationally efficient	False positives are reported. Novel predictions are not possible	http://infernal.janelia.org/	44
It applies RNAFOLD to generate secondary structure from sequence alignments	It can handle alignments with minimum of 10 sequences.	http://groups.csail.mit.edu/cb/MSARi/	42
It is part of sRNA annotation pipeline used in Rfam database. RNAZ can be applied for large scale genomic screens	It requires a fixed sequence alignment as input. Poor sensitivity with low pairwise sequence identity	http://www.tbi.univie.ac.at/~wash/RNAz/	16
sRNA specific promoters, terminator signals were applied to identify IGR sRNAs. It predicts maximum number of known sRNAs in enterobacteriaceae	Current dataset has sensitivity with medium and low %GC genomes	http://cluster.physics.iisc.ernet.in/sRNAscanner/	19
Simple method to predict the sRNA locations with existing information from other databases	Fully depend on the information from other databases. Not possible to work with strains not indexed in other databases.	http://newbio.cs.wisc.edu/sRNA/	18, 50
Continuous atypical positions only considered as possible RNA regions	Not applied for particular RNA family	http://www.rnaspace.org/	52
Functional RNA elements double helices, uridine turns, UNCG loops, tetraloop receptors and mis-pairs are trained. It has high accuracy if motifs added with free energy of folding	Reliability is questionable due to lack of experimental validation	http://rnagene.lbl.gov/	51
Maximally scoring substrings of the input genome above the threshold are identified as RNA regions	Family specific RNA identification is not possible	http://compbio.cs.sfu.ca/nwp-content/software/taverna/	30
First Orthology based method successfully applied to predict sRNA specific gene clusters	Identification of 'novel' sRNAs and flanking genes not having KO numbers are not possible	http://bioserver1.physics.iisc.ernet.in/psrna/	53
Search of 'RNA-rich' cluster in query genomes will identify sRNAs	Search is only restricted with the sRNAs reported in the reference and tracking of 'novel' sRNA is not possible	http://rna.igmors.u-psud.fr/NAPP/	54

tools is highly variable and cannot be compared in the absence of perfect benchmarking data. Each and every approach has its own sensitivity and specificity rates and standard statistical evaluations such as ROC curves[19,57] were not computed for most of these approaches. The ab initio methods applying RNA-specific features have retained most of the known sRNAs based on their preferential occurrence of those features; however, they have also resulted in a large number of 'false positives'. The controlled vocabularies used in the detection of putative sRNA regions might be applicable only with flanking genes assigned with those numbers. Few of the above discussed sRNA finders like ISI were not updated in

recent times. Hence, the users are suggested to select the most recently published or updated tools to identify sRNAs.

In the near future, the complexity of software tools might become condensed and easy usage may be expected. To make further progress in this field a consensus based 'Hybrid' approach which includes the positive aspects of transcriptional signals and consensus secondary structures, together with thermodynamic stability analysis and ncRNA specific features, could give more precise sRNA annotations. Recently, many studies have reported the global transcriptional map of pathogens using RNA-Seq technology;[58–60] these studies have all demonstrated the difficulties involved in the mapping of sRNA reads against the genome of interest. To make valid sRNA annotations, any proposed sRNA identification tool should be equipped to analyze the recent High-throughput data generated from the cDNA microarrays, Next Generation Sequencing (RNA-Seq) experiments, and genomic data. The proposed consensus method should also be available as a graphical user interfaces (GUI) and web server, with a goal of having a greater impact on community annotation.

Conclusions

This review discusses the computational approaches applied in available sRNA prediction tools. Comparative genomics, structure consensus, transcriptional signal-based tools and ab initio protocols were interpreted and their working mechanisms comprehensively analyzed. In the last decade, computational screening has revolutionized the sRNA detection process and become enormously significant. However, experimental validation proves the authenticity and presence of computational predictions of novel regulatory sRNAs. Nevertheless, sRNA prediction employing a less complex and user friendly approach is necessary in the current genomic era and in the near future.

Acknowledgements

The authors thank the University Grants Commission (UGC), Government of India for supporting the Networking Resource Centre in Biological Sciences, Madurai Kamaraj University. JS thanks the Department of Biotechnology, UGC and CSIR, Government of India for funding the research projects.

Author Contributions

JS conceived the idea and wrote the review. PGS contributed to the design of the manuscript and finalized the review. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosure and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Huttenhofer A, Schattner P, Polacek N. Noncoding RNAs: hope or hype? *Trends Genet*. 2005;21:289–97.
2. Johansson J, Cossart P. RNA-mediated control of virulence gene expression in bacterial pathogens. *Trends Microbiol*. 2003;11:280–5.
3. Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z. The microRNA world: small is mighty. *Trends Biochem Sci*. 2003;28:534–40.
4. Papenfort K, Vogel J. Regulatory RNA in Bacterial Pathogens. *Cell Host and Microbe*. 2010;8:116–27.
5. Postic G, Frapy E, Dupuis M, et al. Regulation of virulence in *Francisella tularensis* by small noncoding RNAs. *Nature Preceedings*. 2011. doi:10.1038/npre.2011.5965.1.
6. Toledo-Arana A, Repoila F, Cossart P. Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol*. 2007;10:182–8.
7. Murphy ER, Payne SM. RyhB, an Iron-Responsive Small RNA Molecule, Regulates *Shigella dysenteriae* Virulence. *Infect Immun*. 2007;75:3470–7.
8. Padalon-Brauch G, Hershberg R, Elgrably-Weiss M, et al. Small RNAs encoded within genetic islands of *Salmonella typhimurium* show host-induced expression and role in virulence. *Nucleic Acids Res*. 2008;36:1913–27.
9. Kay E, Reimmann C, Haas D. Small RNAs in Bacterial Cell to Cell Communication. *Microbe*. 2006;1:63–9.
10. Berghoff BA, Glaeser J, Sharma CM, Vogel J, Klug G. Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaerioides*. *Mol Microbiol*. 2009;74:1497–512.
11. Vogel J. A rough guide to the noncoding RNA world of *Salmonella*. *Mol Microbiol*. 2009;71:1–11.
12. Argaman L, Hershberg R, Vogel J, et al. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol*. 2001;11:941–50.

13. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*. 2001;15:1637–51.
14. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*. 2001;2:8.
15. Chen S, Lesnik EA, Hall TA, et al. A Bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*. 2002;65:157–77.
16. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*. 2005;102:2454–9.
17. Pichon C, Felden B. Intergenic sequence inspector: searching and identifying bacterial RNAs. *Nucleic Acids Res*. 2003;19:1707–9.
18. 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.
19. Sridhar J, Narmada SR, Sabarinathan R, et al. sRNAscanner: A Computational Tool for Intergenic Small RNA Detection in Bacterial Genomes. *PLoS ONE*. 2010;5:e11970.
20. Herbig A, Nieselt K. nocoRNAc: Characterization of noncoding RNAs in prokaryotes. *BMC Bioinformatics*. 2011;12:40.
21. Gardner PP, Daub J, Tate JG, et al. Rfam: updates to the RNA families database. *Nucl Acids Res*. 2009;37:D136–40.
22. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2008;36:D25–30.
23. Mercer TR, Dinger ME, Mattick JS. Long noncoding RNAs: insights into functions. *Nat Rev Genet*. 2009;10:155–9.
24. Billoud B, Kontic M, Viari A. Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence databases. *Nucleic Acids Res*. 1996;24:1395–403.
25. Macke T, Ecker D, Gutell R, Gautheret D, Case DA, Sampath R. RNAMotif—A new RNA secondary structure definition and discovery algorithm. *Nucleic Acids Res*. 2001;29:4724–35.
26. Gautheret D, Major F, Cedergren F. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *CABIOS*. 1990:325–11.
27. Gruber AR, Bernhart SH, Zhou Y, Hofacker IL. RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures, Lecture notes in informatics. *Proceedings of German Conference on Bioinformatics* 2010. Braunschweig, Germany.
28. 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.
29. Tjaden B. Prediction of small, noncoding RNAs in bacteria using heterogenous data. *J Math Biol*. 2008;56:183–200.
30. Salari R, Aksay C, Karakoc E, Unrau PJ, Hajirasouliha I, Sahinalp SC. smyRNA: A Novel Ab Initio ncRNA Gene Finder. *PLoS ONE*. 2009;4:e5433.
31. Huttenhofer A, Vogel J. Experimental approaches to identify noncoding RNAs. *Nucl Acids Res*. 2006;34:635–46.
32. Vogel J, Wagner EGH. Target identification of small noncoding RNAs in bacteria. *Curr Opin Microbiol*. 2007;10:262–70.
33. Panek J, Bobek J, Mikulik K, Basler M, Vohradsky J. Biocomputational prediction of small noncoding RNAs in *Streptomyces*. *BMC Genomics*. 2008;9:217.
34. Voß B, Georg J, Schön V, Ude S, Hess WR. Biocomputational prediction of noncoding RNAs in model cyanobacteria. *BMC Genomics*. 2009; 10:123.
35. Schluter JP, Reinkensmeier J, Daschkey S, et al. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. *BMC Genomics*. 2010;11:245.
36. Postic G, Frapy E, Dupuis M, et al. Identification of small RNAs in *Francisella tularensis*. *BMC Genomics*. 2010;11:625.
37. Liang H, Zhao YT, Zhang JQ, Wang XJ, Fang RX, Jia YT. Identification and functional characterization of small noncoding RNAs in *Xanthomonas oryzae* pathovar *oryzae*. *BMC Genomics*. 2011;12:87.
38. Chen Y, Indurthi DC, Jones SW, Papoutsakis ET. Small RNAs in the Genus *Clostridium*. *mBio*. 2011;2:e00340–e00310.
39. Gautheret D, Lambert AA. Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. *J Mol Bio*. 2001;313:1003–11.
40. Lorenz AR, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA Websuite. *Nucleic Acids Res*. 2008;36:W70–4.
41. Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994;22:4673–80.
42. Coventry A, Kleitman DJ, Berger B. MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc Natl Acad Sci U S A*. 2004;101:12102–7.
43. Griffiths-Jones S, Moxon R, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating noncoding RNAs in complete genomes. *Nucleic Acids Res*. 2005;33:D121–4.
44. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009;25:1335–7.
45. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*. 1999;288:911–40.
46. Bernhart SH, Hofacker IL, Will S, Gruber A, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*. 2008;9:474.
47. Touzet H, Perriquet O. CARNAC: folding families of related RNAs. *Nucleic Acids Res*. 2004;32:W142–5.
48. Kingsford C, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcriptional terminators illuminates their relationship to DNA uptake. *Genome Biol*. 2007;8:R22.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
50. Livny J, Teonadi H, Livny M, Waldor MK. High-Throughput, kingdom wide prediction and annotation of bacterial noncoding RNAs. *PLoS One*. 2008;3:e3197.
51. 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.
52. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A*. 2002;99:7542–7.
53. Sridhar J, Rafi ZA. Small RNA identification in *Enterobacteriaceae* using synteny and genomic backbone retention. *OMICS*. 2007;11:74–99.
54. Ott A, Idali A, Marchais A, Gautheret D. NAPP: the nucleic acid phylogenetic profile database. *Nucl Acids Res*. 2011. doi:10.1093/nar/gkr807.
55. Sridhar J, Sowmya G, Sekar K, Rafi ZA. PsRNA: A Computing Engine for the comparative identification of putative small RNA locations within intergenic regions. *Genomics Proteomics Bioinformatics*. 2010;8: 127–34.
56. Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics*. 2008;24:2807–13.
57. Lu X, Goodrich-Blair H, Tjaden B. Assessing computational tools for the discovery of small RNA genes in bacteria. *RNA*. 2011;17:1635–47.
58. Sharma CM, Hoffman S, Darfeuille F, et al. The primary transcriptome of the major human pathogen. *Helicobacter pylori*. *Nature*. 2010;464: 250–5.
59. Kumar R, Lawrence ML, Watt J, Cooksey AM, Burgess SC, Nanduri B. RNA-Seq based transcriptional map of bovine respiratory disease pathogen "Histophilu somni 2336". *PLoS One*. 2012;7(1):e0029435.
60. Pellin D, Miotto P, Ambrosi A, Cirillo DM, Serio CD. A genomewide indentification anlaysis of small regulatory RNAs in Mycobacterium tuberculosis by RNA-seq and conservation analysis. *PLoS One*. 2012;7(3): e0032733.