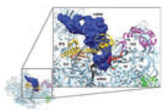
RNA Biology



ISSN: 1547-6286 (Print) 1555-8584 (Online) Journal homepage: https://www.tandfonline.com/loi/krnb20

A comparative study of sequence- and structure-based features of small RNAs and other RNAs of bacteria

Amita Barik & Santasabuj Das

To cite this article: Amita Barik & Santasabuj Das (2018) A comparative study of sequence- and structure-based features of small RNAs and other RNAs of bacteria, RNA Biology, 15:1, 95-103, DOI: 10.1080/15476286.2017.1387709

To link to this article: https://doi.org/10.1080/15476286.2017.1387709

RNA BIOLOGY
2018, VOL. 15, NO. 1, 95–103
https://doi.org/10.1080/15476286.2017.1387709

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

A comparative study of sequence- and structure-based features of small RNAs and other RNAs of bacteria

Amita Barik[a] and Santasabuj Das[a,b]

[a]Biomedical Informatics Centre, National Institute of Cholera and Enteric Diseases, Kolkata, West Bengal, India; [b]Division of Clinical Medicine, National Institute of Cholera and Enteric Diseases, Kolkata, West Bengal, India

ABSTRACT

Small RNAs (sRNAs) in bacteria have emerged as key players in transcriptional and post-transcriptional regulation of gene expression. Here, we present a statistical analysis of different sequence- and structure-related features of bacterial sRNAs to identify the descriptors that could discriminate sRNAs from other bacterial RNAs. We investigated a comprehensive and heterogeneous collection of 816 sRNAs, identified by northern blotting across 33 bacterial species and compared their various features with other classes of bacterial RNAs, such as tRNAs, rRNAs and mRNAs. We observed that sRNAs differed significantly from the rest with respect to G+C composition, normalized minimum free energy of folding, motif frequency and several RNA-folding parameters like base-pairing propensity, Shannon entropy and base-pair distance. Based on the selected features, we developed a predictive model using Random Forests (RF) method to classify the above four classes of RNAs. Our model displayed an overall predictive accuracy of 89.5%. These findings would help to differentiate bacterial sRNAs from other RNAs and further promote prediction of novel sRNAs in different bacterial species.

ARTICLE HISTORY
Received 1 June 2017
Revised 26 September 2017
Accepted 28 September 2017

KEYWORDS
Base-pairing propensity; base-pair distance; motifs; minimum free energy of folding; random forests; small RNAs; shannon entropy

Introduction

Traditionally, RNA is considered as a genetic carrier for transfer of information from DNA to proteins. It is not surprising that most studies have dealt with messenger RNAs (mRNAs), ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), which are associated with the cellular machinery for protein translation. However, a large number of RNAs with multitude of regulatory functions, which were independent of protein translation, were discovered lately across many kingdoms, from plants to humans. Regulatory RNAs in bacteria are a heterogenous group of molecules that modulate almost every metabolic function of the cells. Examples of bacterial RNA regulators include riboswitches, small RNAs and CRISPR (clustered regularly interspaced short palindromic repeats) RNAs (reviewed in [1]). The present study is focused on small RNAs (sRNAs), which are generally untranslated (with few exceptions), encoded in the 'empty' intergenic regions of bacterial chromosomes[2,3] and reported to perform a myriad of molecular functions regulating bacterial physiology, cell envelope architecture, intermediate metabolism, cell-cell communication, biofilm formation and virulence.[1,4,5] Bacterial sRNAs are shown to exert their role by base-pairing with the target mRNAs or by modulating protein functions.[1,5,6] However, bifunctional sRNAs with roles in translation and post-transcriptional regulation have also been documented. RNAIII of *Staphylococcus aureus*, apart from binding to mRNAs, which are translated into virulence factors, encodes a 26 amino acid δ-hemolysin peptide as well.[7] Similarly, *Escherichia coli* SgrS RNA that blocks translation of ptsG mRNA into a sugar-phosphate transporter, is translated to a 43 amino acid protein, SgrT.[8]

Small RNAs are generally classified into three functional groups, based on their target genes: (1) *cis-encoded* sRNAs that originate from the same genetic locus as the target mRNAs and bind extensively to them (2) *trans-encoded* sRNAs that are encoded by genes located at different genetic loci and share limited complementarity with targets, and (3) protein-binding *sRNAs* that regulate target genes after binding to proteins instead of mRNAs.[9–11] Several *trans-encoded* sRNAs in Gram negative bacteria depend on the RNA chaperone protein, Hfq for efficient base-pairing with their target mRNAs.[12–16] In fact, Hfq co-immunoprecipitation with sRNAs has been successful to identify novel sRNAs in many bacteria, including *E. coli*[17] and *Salmonella enterica* serovar Typhimurium.[18] Majority of the studies on sRNA-mediated regulation in the past were carried out with Gram negative bacteria.[19,20] As a result, sRNAs analysed in the present study are mostly derived (70%) from these organisms. We, however, also compared the sequence and structural features of sRNAs from Gram positive and Gram negative bacteria to find out the difference between them, if any.

Breakdown products of mRNA transcripts as well as other RNAs, such as tRNAs and rRNAs might contaminate sRNA samples isolated from cytoplasmic total RNA extracts. For accurate computational prediction of sRNAs, it is important to

CONTACT Dr. Santasabuj Das dasss.niced@mail.gov.in, santasabujdas@yahoo.com ICMR-National Institute of Cholera and Enteric Diseases, P-33, C. I. T. Road Scheme XM, Beliaghata, Kolkata - 700010, West Bengal India.

Supplemental data for this article can be accessed on the publisher's website.

discriminate between different RNA categories.[21] The last decade has accumulated a wealth of information and many computational algorithms have been developed for successful screening of sRNAs.[22–26] Unlike protein-coding genes, which exhibit strong ORFs, codon bias signals and ribosome-binding (Shine Dalgarno) sequences, most sRNAs are poorly conserved and generally, do not share common statistical patterns.[6,27,28] However, comparative sequence analysis and thermodynamic stability were applied to identify sRNAs in most sequenced bacterial genomes and these approaches appear promising.[25,29,30] Given that the methods developed were mostly applied to certain specific bacterial strains, there is an urgent need to identify sRNAs features, which could be universally exploited to develop a robust and accurate prediction algorithm applicable to a large number of bacteria.

We investigated the sequence- and structure-based features that could be useful to differentiate sRNAs from other RNAs such as tRNAs, rRNAs and mRNAs. We curated a heterogeneous collection of experimentally validated bacterial sRNAs from Bacterial Small Regulatory RNA Database (BSRD)[10] and compared them with tRNAs, rRNAs and mRNAs. We found that sRNAs differed from other RNAs with respect to a number of features, including length, G+C composition and minimum free energy of folding. Although, functions of RNAs might depend on their "secondary structures" like hairpins and stem-loops, these are not unique to sRNAs. We looked into the presence of tetraloop motifs in different classes of RNAs and observed that motif frequencies might be used as a reliable feature to differentiate sRNAs from other RNAs. It was earlier reported that micro-RNAs (miRNAs) in eukaryotes have distinct RNA-folding measures, such as Shannon entropy, base-pair distance and base-pairing propensity.[21,31] We found that these folding measures could distinguish sRNAs from other bacterial RNAs. Finally, a predictive model was developed based on the above features and using Random Forests (RF) method. The model classified the four different categories of bacterial RNAs (sRNAs, tRNAs, rRNAs and mRNAs) with an overall predictive accuracy of 89.5%. The features identified in the present work could be leveraged to design computational algorithms to explore novel sRNAs in many other bacteria.

Results

Variation in RNAs count and length

The non-redundant sRNA dataset curated from BSRD[10] comprises of 816 sRNAs isolated from 33 different bacterial species (Table S1). Lengths of different RNAs present in our dataset are shown in Table 1. Majority of the sRNAs in our dataset are from Gram negative bacteria, such as *Pseudomonas* (150), *Synecocystis* (75) and *Helicobacter* (70) (the number in the parentheses represents the number of sRNAs identified in that bacterial species). We followed the same classification as BSRD for sRNAs. Based on the target molecules, sRNAs are classified in BSRD into two major categories: (1) mRNA-binding antisense sRNAs and (2) protein-binding sRNAs. mRNA-binding sRNAs are further subdivided into *cis-encoded* and *trans-encoded* sRNAs. We classified the sRNAs in our dataset into three classes: *cis, trans* and *miscellaneous* (*misc*). The *miscellaneous* category comprises of the protein-binding sRNAs as well as those sRNAs, which were mentioned as "regulatory element" in the BSRD database and did not belong to *cis* or *trans* category. Of 816 sRNAs in our dataset, 189 (23%) are *cis-encoded*, 596 (73%) are *trans-encoded* and 31 (4%) come under the *miscellaneous* category.

The variation in the size of sRNAs in our dataset is extreme, ranging from 29 nt to 1501 nt. While in *cis-encoded* sRNAs, the length varies from 29 nt (*Synecocystis*) to 1501 nt (*Pseudomonas*), the size of *trans-encoded* sRNAs ranges between 34 nt (*Helicobacter*) and 1363 nt (*Staphylococcus*). Under the *miscellaneous* category, length of sRNAs ranges from 37 nt (*Mycobacterium*) to 838 nt (*Staphylococcus*). The average lengths of sRNAs in *cis, trans* and *misc.* category are 219 nt, 182 nt and 296 nt, respectively. In the tRNA dataset, the smallest tRNA is 68 nt that belongs to *Xanthomonas*, while the longest one is of *Myxococcus* with a length of 99 nt. In case of rRNAs, length ranges between 1200 nt (*Acinetobacter*) and 3147 nt (*Burkholderia*). The mRNAs in our dataset also display a wide variation in length, ranging from 66 nt (*Escherichia* and *Salmonella*) to 4245 nt (*Bordetella*). The range of length of sRNAs present in our dataset is shown in Fig. 1. The histogram illustrates the large variation in length of sRNAs with a predominant peak at 51–150 nt, followed by a second peak at 151–250 nt. Analysis of size distribution of the sRNAs in the dataset shows that most of them (76%) are 50 nt to 250 nt long. In terms of length, sRNAs belong to the intermediate category, being longer than tRNAs, but shorter compared to rRNAs and mRNAs. However, few sRNAs (5%) in our dataset are longer than 500 nt. These include 13 sequences from *cis*, 23 from *trans* and 3 from *miscellaneous* category. Most of these sRNAs belong to *Pseudomonas* and *Staphylococcus* and have lengths ranging from 500–1500 nt.

G+C composition of RNAs

The average G+C composition of different categories of RNAs is shown in Table 1. We observed that tRNAs, rRNAs and mRNAs have relatively higher G+C content (58.9±4.7, 53.0±1.8 and 52.8±12.9, respectively) as compared to sRNAs, which have the average G+C content of 49.9±13.0. The P value (at $\alpha = 0.05$ level of significance) obtained from the Analysis of Variance (ANOVA) test on the G+C content is 1.0507E-100, showing significant statistical difference in the mean values amongst the four classes of RNAs. Out of the sRNA subclasses, the *miscellaneous* category has a similar G+C composition (average ~58.5±9.4) like tRNAs.

RNA motifs

The frequency of five different motifs (please refer to Materials and Methods) was calculated in all the four classes of RNAs. Table 2 shows the average frequency of the five motifs in sRNA, tRNA, rRNA and mRNA. The P value calculated by ANOVA is less than α (0.05), rejecting the null hypothesis that the occurrence of the motifs is similar in different classes of RNAs. We therefore chose these motif frequencies as potential features to train our classifier. We observed that three members of the GNRA family (GCAA, GAAA and GAGA) are very

Table 1. Statistics for different RNA sequences.

Parameters	tRNA[a]	rRNA[b]	mRNA[c]	sRNA[d]			
				Cis	*Trans*	*Misc.*	All
No. of RNAs	1421	271	990	189	596	31	816
Range of Length (nt)	68–99	1200–3147	66–4245	29–1501	34–1363	37–838	29–1501
Avg Length (nt)	77±5	1735±587	971±614	219±180	182±155	296±164	195±164
Avg (G+C)%	58.9±4.7	53.0±1.8	52.8±12.9	50.9±10.5	49.1±13.7	58.5±9.4	49.9±13.0
MFE (kcal/mol)	−30.48±4.45	−628.64±217.10	−348.94±253.76	−73.16±75.8	−58.04±51.88	−109.75±56.42	−63.51±59.42
AMFE (kcal/mol)	39.18±4.5	36.20±1.6	35.14±10.6	30.38±10.5	32.65±11.4	37.78±8.3	32.32±11.1
MFEI	0.67±0.06	0.68±0.02	0.65±0.06	0.59±0.14	0.67±0.14	0.65±0.10	0.65±0.14
Nbp	0.31±0.03	0.32±0.01	0.32±0.03	0.29±0.05	0.29±0.05	0.31±0.02	0.29±0.05
NQ	0.20±0.10	0.37±0.08	0.40±0.11	0.28±0.14	0.23±0.14	0.29±0.12	0.24±0.14
ND	0.07±0.04	0.11±0.02	0.12±0.03	0.09±0.04	0.07±0.04	0.10±0.04	0.08±0.04

Values are represented as the mean ± standard deviation.
[a]Transfer RNAs (tRNAs) of 32 bacterial strains were collected from tRNAscan-SE Genomic tRNA Database.
[b]Ribosomal RNAs (rRNAs) of both large (LSU) and small (SSU) subunits of the 33 bacterial strains were collected from SILVA database.
[c]Messenger RNAs (mRNAs) of 33 bacterial strains were randomly collected from NCBI genome database.
[d]Experimentally verified bacterial small RNAs (sRNAs) from 33 different bacterial strains were curated from BSRD.

common in rRNAs, while in tRNAs, the UUCG motif is most predominant. These consensus sequences are also widely observed in sRNAs and their average frequency is similar to that occurring in mRNAs. We found that occurrence frequency of the motif CUUG in sRNAs is high compared to other three classes of RNAs.

RNA folding measures

The average minimum free energy (MFE), Adjusted MFE (AMFE) and minimal folding free energy index (MFEI) (Refer Materials and Methods) of different classes of RNAs are reported in Table 1. In our dataset of 816 sRNAs, the negative free energies of folding ranges from −1.7 to −614.8 kcal/mol with an average of −63.51 kcal/mol. A majority of them (80%) have a negative MFE of 10–50 kcal/mol. MFE and length are strongly correlated in all the four classes with correlation coefficient of −0.63, −0.98, −0.89 and −0.89 for tRNA, rRNA, mRNA and sRNA, respectively. When the normalized MFE (AMFE) and MFEI were studied in different classes, it was observed that an average sRNA has a significantly lower AMFE and MFEI than an average tRNA and rRNA (P value < 0.05). However, it is difficult to distinguish sRNAs from mRNAs, since they have similar AMFE and MFEI range and almost similar average values. AMFE and MFEI of more than 80% of sRNAs fall in the range of corresponding values for mRNAs (10 to 50 kcal/mol and 0.45 to 0.75, respectively).

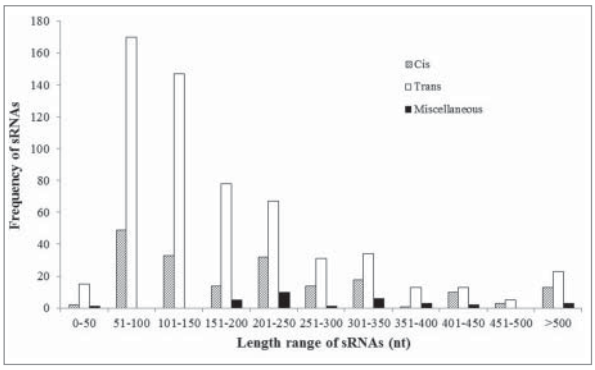


Figure 1. Length distribution of the different types of sRNAs. Majority of the sRNAs have a length ranging from 50–250 nt.

Figure 2 depicts the distribution of Nbp, NQ and ND in the four different classes of RNAs. Nbp usually ranges from 0 to 0.5; 0 for no base-pair interactions and 0.5 for all base-pairings. In all classes of RNAs, the variation of Nbp is generally between 0.24 and 0.42 with a broad peak at 0.36, thus confirming the presence of loops in the secondary structure of all RNAs. In case of sRNA and tRNA dataset, the majority of the sequences have a NQ value in the range of 0.22–0.32, while for rRNAs and mRNAs, NQ shows a broad peak from 0.33 to 0.55. Similar trend is observed for ND, with both sRNAs and tRNAs having a wide range from 0.036 to 0.107, while rRNAs and mRNAs show a range of ND largely between 0.107 and 0.178.

sRNAs in Gram negative and Gram positive bacteria

The sRNA dataset is classified based on the type of bacteria from which the RNA sequence was isolated. Of the 816 sRNAs, 573 (70%) belong to Gram negative bacteria, while the rest 243 (30%) fall under the category of Gram positive bacteria. The average G+C content of sRNAs in Gram negative bacteria is 52% and that of Gram positive bacteria is 44%. Usually sRNAs in Gram negative bacteria bind to Hfq chaperone protein to carry out different physiological processes, including stress resistance and virulence. In our dataset, we have 68 *trans-encoded* sRNAs that are reported to bind Hfq chaperone protein. Of these, only three sRNAs (LhrC-2, LhrC-4 and LhrC-5) belong to Gram positive bacteria (*Listeria monocytogenes* EGD-e) (Table 3). These Hfq binding sRNAs have an average length

Table 2. The average frequency of five different motifs present in four classes of RNAs. P value reported from ANOVA in each case is shown.

Average motif frequency	Classes of RNAs				P value ($\alpha = 0.05$)
	sRNA	tRNA	rRNA	mRNA	
UUCG	0.35	1.18	0.34	0.38	0
GCAA	0.41	0.27	0.54	0.53	9.42E-40
GAAA	0.56	0.20	0.59	0.59	3.03E-80
GAGA	0.28	0.21	0.54	0.31	3.05E-28
CUUG	0.37	0.29	0.35	0.31	0.001613

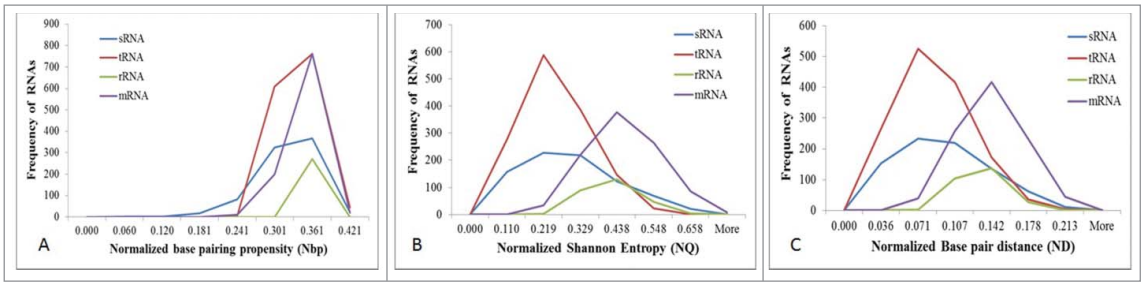


Figure 2. Distribution of (A) normalized base-pairing propensity (Nbp), (B) normalized Shannon entropy (NQ) and (C) normalized base-pair distance (ND) in different classes of RNAs.

of 118 nt and analysis of their base composition reveals that they tend to display an elevated percentage of A+U content (mean ~55%). The average Nbp of these sRNA is 0.27, which suggests the presence of more loops in their secondary structures. However, these sRNAs have a lower NQ (mean ~0.18) and ND (mean ~0.06) than an average sRNA (mean NQ ~0.24, ND ~0.08), indicating a well-defined structure. GCAA and GAAA motifs of the GNRA family and the UUCG motif are abundant in the sRNAs that bound to Hfq protein, though there are variations among different bacterial species.

Performance of random forest model

The Random Forest Model generated by WEKA (Waikato Environment for Knowledge Analysis)[32] Version 3.4.3 (Refer Materials and Methods) correctly classifies 89.2% (2688/3015) of the instances in the Training dataset (Table 4).

In the training dataset of 3015 instances, the model correctly predicts 74%, 97%, 97% and 86% instances of sRNAs, tRNAs, rRNAs and mRNAs, respectively. Figure 3A depicts the confusion matrix of the model on the training dataset. Re-evaluation of the RF model on a test dataset of 754 instances results in correct classification of 92% instances. The model correctly identifies 77%, 99%, 97% and 91% of the instances in the test dataset of sRNA, tRNA, rRNA and mRNA classes, respectively (Fig. 3B).

WEKA enlists the performance measures (please refer to the footnote of Table 5 for details) of the developed model for each class as well as the overall performance. Table 5 reports only the final weighted values of the different performance measures in case of training and test datasets. The final value calculated by WEKA is the weighted average where the number of instances in each class defines the weighs. For example, if *n1, n2, n3, n4* represent the numbers of instances and *TP1, TP2, TP3, TP4* represent the True Positive (TP) rates of the four classes, then the final weighted TP rate will be $(n1^{*}TP1 + n2^{*}TP2 + n3^{*}TP3 + n4^{*}TP4)/(n1+n2+n3+n4)$. In a similar way, the weighted averages of other performance measures were calculated in WEKA. Both the TP rate and precision value of the RF classifier on training dataset is 0.89. The classifier achieves a TP rate of 0.92 and precision value of 0.92 on test dataset, explaining the robustness of the model.

Discussion

Small RNAs form an important functional layer in bacteria. These RNAs usually do not encode proteins, but they regulate a plethora of biological processes in bacteria. Identification and functional characterization of sRNAs are essential for understanding the sRNA-mediated regulatory networks in bacteria. To complement the experimental methods, many computational approaches have been developed for identifying sRNAs in different bacterial species. However, computational detection of sRNAs remains extremely challenging, since very few signatures specific to them have been described.

The present work is an endeavour to enlist different sequence and structural features of bacterial sRNAs, which could differentiate them from other RNAs. We investigated a comprehensive and heterogeneous collection of experimentally verified 816 sRNAs curated from the BSRD database and compared them with rRNAs, tRNAs and mRNAs.

The length of sRNAs in our dataset varies from 29–1501 nts, with the mean value of 195 nt. Our results are in good agreement with previous studies, where the average length of a typical bacterial sRNA was reported to be 200–250 nt.[2,6] While rRNAs are much longer (mean ~1735 nt) and have more complex secondary and tertiary structure,[33] tRNAs are the shortest RNAs (mean length ~77 nt). The average length of 990 mRNA sequences was found to be 971 nt. Although, there is a wide variation in the length of sRNAs, they differ strikingly in length

Table 3. Average measures of sRNAs binding to Hfq chaperone protein in different bacterial species.

Bacteria species	No. of sRNAs binding to Hfq	Length	GC%	UUCG	GCAA	GAAA	GAGA	CUUG
Escherichia	19	133	45.9	0.57	0.39	0.63	0.37	0.25
**Listeria*	3	112	36.5	0.00	0.60	0.59	0.00	0.00
Pseudomonas	2	143	50.9	0.47	0.94	0.00	0.00	0.00
Salmonella	34	110	45.9	0.45	0.47	0.50	0.19	0.18
Vibrio	9	105	42.9	0.10	0.12	0.10	0.00	0.41
Yersinia	1	206	42.7	0.00	0.49	0.97	0.00	0.00

**Listeria* is the only Gram positive bacteria in our dataset that has sRNAs reported to bind to Hfq protein.

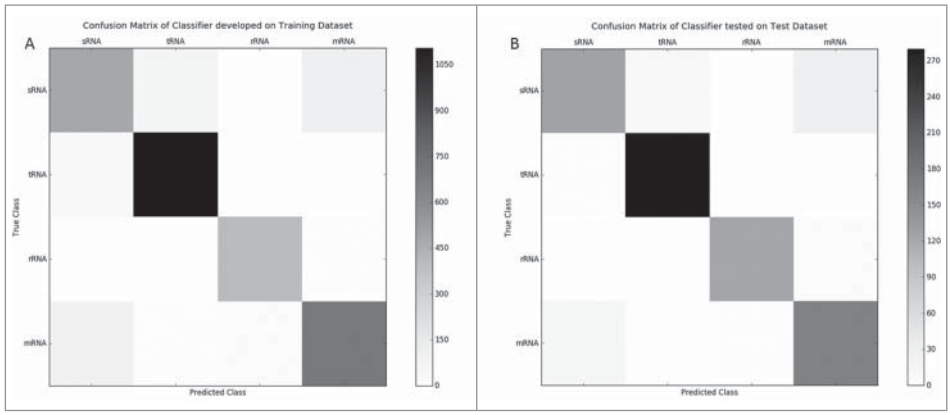


Figure 3. Confusion matrix of the Random Forest model on (A) training dataset and (B) test dataset. The diagonal in the matrices represents the correctly identified instances.

from rRNAs, tRNAs and mRNAs. The *trans-encoded* sRNAs (mean length ~182 nt) are typically shorter than the *cis-encoded* sRNAs (mean length ~219 nt). A comparison between different classes of sRNAs reveals that there are significant differences in terms of length, G+C content, AMFE, MFEI, Nbp, NQ and ND between the classes (P value < 0.05). In a separate study, Vazquez-Anderson and Contreras[34] have highlighted the sequence and structural differences between the *cis* and *trans* sRNAs. However, during the development of the classifier to distinguish sRNAs from other RNAs, we considered all the sRNAs under one class.

Sequence content statistics, such as base composition (especially G+C composition) have been widely employed in computational algorithms to identify novel sRNAs in bacteria.[26,35,36] Since G:C and A:U form three and two hydrogen bonds (H-bonds), respectively, a high G+C content in RNA makes them more stable. Previous studies have also reported that G+C content of structural RNA has a strong correlation with the optimal temperature (T_{opt}) for growth of living organisms.[37,38] In another study, Galtier and Lobry[39] observed that genomic G+C content and T_{opt} did not display good correlations, although they were strongly correlated for tRNAs and rRNAs. G+C content, thus, have a strong impact on the stability of RNA due to three H-bonds. Our results confirmed this observation where we found that tRNAs being more structured, showed a higher G+C content (58%) than the rest of the RNAs. sRNAs, on an average, have nearly equal G+C and A+U contents. However, this varies in different classes of sRNAs. It is intriguing that sRNAs under the *miscellaneous* category have a higher G+C content (58.5±9.4) than that of *cis-encoded* (50.9±10.5) and *trans-encoded* (49.1±13.7) sRNAs. In fact, the average G+C content of the *miscellaneous* group of sRNAs is closer to that of tRNAs (59%) and rRNAs (53%), which suggests that the former class of sRNAs tend to form a rigid structure. The relatively low G+C contents of *cis-* and *trans-encoded* sRNAs makes them less stable and this could be the apparent reason why they tend to favour RNA-RNA interactions, forming stable duplexes.

The average G+C content of whole genomes in most organisms (82%) are not considerably different from sRNAs and

Table 4. Statistics of data classified by the RF Model in different classes of RNA.

Class	sRNA	tRNA	rRNA	mRNA	All
Total no. of instances	816	1421	542	990	3769
No. of instances in Training Dataset (80%)	654	1138	415	808	3015
Correctly identified instances	481	1105	403	699	2688
Incorrectly identified instances	173	33	12	109	327
No. of instances in Test Dataset (20%)	162	283	127	182	754
Correctly identified instances	126	280	123	165	694
Incorrectly identified instances	36	3	4	17	60

Table 5. Performance measures of the RF classifier on Training and Test dataset.

Parameters	On Training Dataset (3015 instances)	On Test Dataset (754 instances)
[a]TP Rate (Sensitivity or Recall)	0.89	0.92
[b]FP Rate	0.04	0.03
[c]Precision	0.89	0.92
[d]F-measure	0.89	0.92
[e]ROC-area	0.97	0.98

[a]True Positive (TP) rate or Recall is the proportion of instances which were classified as class *x*, among all instances which truly belong to class *x*.
[b]False Positive (FP) rate is the proportion of instances which were classified as class *x*, but belong to a different class, among all instances which are not of class *x*.
[c]Precision is the proportion of the instances which truly have class *x* among all those which were classified as class *x*.
[d]The F-Measure is 2 * Precision * Recall / (Precision + Recall), a combined measure for precision and recall.
[e]Receiver operating characteristic (ROC) curve is a graphical plot that is created by plotting the true positive rate (TPR) against the false positive rate (FPR).

mRNAs. However, tRNAs and rRNAs in these cases display an elevated G+C content owing to their rigid structure. Analysis of the base composition in whole genome sequences of Gram positive and Gram negative bacteria shows that the later have an elevated G+C content (45% vs 54%). sRNAs in these bacteria also follow a similar trend: the average G+C content for Gram negative bacteria is 52% as opposed to 44% in Gram positive organisms (P value < 0.05).

Earlier studies have reported that sRNAs in Gram negative bacteria usually require Hfq chaperone to mediate base-pairing with their target mRNAs.[5,12–14,40] Hfq is a hexameric RNA-binding protein that acts as a general cofactor for *trans-encoded* sRNAs. Since the genes for *trans-encoded* sRNAs are located at loci different from those encoding their target RNAs, the sRNA-target complementarity is incomplete, resulting in imperfect duplexes.[41] This could be the apparent reason why Hfq protein assists *trans-encoded* sRNAs in duplex formation and subsequently enhances the RNA-RNA interactions by acting as an RNA chaperone.

Gram positive bacteria generally lack the gene encoding Hfq chaperone protein.[42] However, Hfq protein is reported in few Gram positive organisms. We have 68 *trans-encoded* sRNAs in our dataset, which are reported to bind to Hfq chaperone protein. Among them, only 3 sRNAs belong to Gram positive bacteria (*Listeria monocytogenes* EGD-e strain).[42,43] Analysis of the base composition of 68 Hfq-binding sRNAs revealed that they have low G+C content (mean ~45%). It is conceivable that this might result in a less stable conformation of these RNAs, and hence they might require the assistance of Hfq for binding to their targets. However, examples are also available where sRNAs with low G+C content did not require the assistance of Hfq proteins. For example, in *S. aureus* (32.8% G+C), Hfq appears to be dispensable. Hence, linking G+C content with the requirement for Hfq cannot be generalized. Nevertheless, it is well established that Hfq chaperone prefers to bind AU rich RNA sequences.[44,45]

Comparative sequence analysis has revealed that tetraloops are most common RNA motifs and are involved in various functional and structural roles. Importantly, they are thermodynamically stable with higher melting temperatures and play a vital role in RNA folding, RNA-RNA tertiary interactions and as protein binding sites.[46–48] We studied the occurrence of five common motifs (GCAA, GAAA, GAGA, UUCG and CUUG) in different classes of RNA. In the UNCG family, the UUCG tetraloop is considerably more stable owing to the 2′-OH group.[48,49] We found that tRNAs have a higher frequency of UUCG motif compared with sRNAs, rRNAs and mRNAs (Table 2). GNRA loops are predominant in rRNAs (mean ~0.56), followed by mRNAs and sRNAs (mean ~0.48 and 0.42, respectively), while their frequency is low in tRNAs (mean ~0.23) (P value < 0.05). In GNRA loops, the first and the last nucleotides are involved in G-A base pairing, which in turn is surrounded by a network of heterogeneous H-bonds, thus providing a stability to the RNA secondary structure.[46] Previous studies have reported the abundance of these tetraloops in both rRNAs and tRNAs.[47] In this study, we observed that these motifs are also common in sRNAs, although the frequency of their occurrence is relatively lower than rRNAs and tRNAs. CUUG RNA hairpin, which is thermodynamically stabilized by the intraloop C-G base pair, is mostly found in sRNAs (Table 2).

The knowledge of base pairings in an RNA sequence is critical for understanding its function. We studied three different measures, such as Nbp, NQ and ND in four different classes of RNAs. The measures, NQ and ND have been used to find out if a sequence folded into a unique secondary structure or into several alternative structures.[50] They are used as a measure to indicate the well-definedness of RNA structure. A low value of NQ in a RNA sequence suggests that it has a well-defined structure. Among the four different classes of RNAs, mRNAs and rRNAs display a relatively higher NQ (0.40 and 0.37, respectively) and ND (0.12 and 0.11, respectively). The high value of NQ and ND shown by these RNAs can be attributed to their long length.[51] Though sRNAs have a lower NQ and ND than rRNAs, they possess statistically higher NQ (0.24 ± 0.14) and ND (0.08 ± 0.04) compared with tRNAs (0.20 ± 0.10 and 0.07 ± 0.04, respectively) (P value < 0.05). A relative high NQ and ND values suggests that sRNAs are unstructured and might fold into several alternative structures.

We also calculated the minimum folding energy (MFE) of different RNA sequences. However, MFE of RNA is strongly and positively correlated with its sequence length. Long RNA sequences tend to have a high degree of freedom and fold into complex secondary structures with high thermodynamic stability or low MFE.[21, 33] We normalized MFE with the length and the Adjusted MFE (AMFE) and minimal folding free energy index (MFEI) of RNAs were calculated and used as a comparable measure. Interestingly, both AMFE and MFEI showed variations between different classes of RNAs and might be used as important features to distinguish sRNAs from other RNAs in the future.

Although length could be one parameter to distinguish different RNAs, we did not consider it in building the classifier, since most of the features were normalized against the length. We used the features G+C content, AMFE, Nbp, NQ and motif frequencies to develop a Random Forest Model that had the potential to be used to classify the four different classes of RNAs. Our model achieved an overall prediction accuracy of 89% and correctly predicts majority of the sRNAs (74%), tRNAs (97%), rRNAs (97%) and mRNAs (86%). The vast heterogeneity shown by sRNAs makes it difficult to classify them accurately and hence, 173 sRNAs were misclassified (63 as tRNAs, 3 as rRNAs and 107 as mRNAs). The RF classifier applied to a blind dataset of 754 instances showed a high TP rate (0.92) and precision value (0.92) and identified 77% of the sRNAs correctly. In spite of the great diversity shown by sRNAs, our model displays high prediction efficiency. The parameters studied in this work may be used to understand bacterial sRNAs in a better way. We believe that the features described in our study would be helpful for researchers to develop computational algorithms, which would be both rapid and efficient to identify novel sRNAs in other bacteria.

Conclusion

This study provides a relative comparison between different classes of RNAs in 33 bacterial species. We used a

comprehensive collection of non-redundant datasets of RNAs. The sRNA dataset was curated from BSRD that contains the most recent and experimentally validated sequences isolated from different bacteria. We observed that the sRNAs in Gram negative bacteria display an elevated G+C content compared with those from Gram positive bacteria. In contrast, sRNAs that bound to Hfq proteins tend to have a lesser G+C content. A number of sequence- and structure-based features of different RNAs were calculated and compared to find out the ones that could most reliably distinguish sRNAs from other classes of RNAs. Based on some selected features, such as G+C content, normalized minimum free energy of folding, normalized base-pairing propensity, normalized Shannon entropy and motif frequencies, a Random Forest Model was developed to classify different classes of RNAs. We found that sRNAs are significantly different from other RNAs with respect to the above features. The above findings will enhance our knowledge about sRNAs and help researchers to develop more accurate and reliable computational algorithms for prediction of novel sRNAs in other bacteria.

Materials and methods

Data curation

For bacterial small RNA dataset, BSRD was searched and all the sRNA sequences that were experimentally identified by northern blotting were collected. We retrieved 897 sRNAs belonging to 33 bacteria, of which 25 were Gram negatives while 8 were Gram positive bacteria. For ribosomal RNA (rRNA) and transfer RNA (tRNA) dataset, sequences of all strains of the above mentioned 33 bacteria were collected from SILVA database[52] and tRNAscan-SE Genomic tRNA database,[53] respectively. We obtained 4299 and 8559 sequences of the large subunit (LSU) and small subunit (SSU) of ribosomal RNA, respectively. In case of tRNA, a total of 5600 sequences from 32 bacterial strains (except *Azotobacter*) were collected. For mRNA dataset, we randomly selected 30 CDS sequences of each of the 33 bacterial strains from National Center for Biotechnology Information (NCBI)[54] genome database. Our mRNA dataset thus, consists of 990 mRNAs.

The sequences in each dataset were clustered using CD-HIT Est[55] server to remove redundancy at 95% sequence similarity. This resulted in a dataset of 817 sRNAs. Of these, one regulatory element isolated from *Pseudomonas aeruginosa* PAO1 (BSRD ID: spae3890.1)[56] was an exceptional case of length 10201 nt and the authors have classified it under the 5′-untranslated region (UTR) category. We excluded this RNA from our dataset and the final sRNA dataset thus comprised of 816 sequences. The final non-redundant dataset of rRNA (both LSU and SSU) and tRNA contained 271 and 1421 sequences, respectively.

Base composition statistics and structural parameters

Previous studies have reported that bacterial sRNAs differ from other RNAs in their base composition.[26,57,58] We calculated the G+C content of different RNA sequences and normalized it against their respective lengths. The RNAfold program from Vienna RNA package[59,60] was used to calculate MFE of the RNAs. The program also gives the most favourable and stable secondary structure of the RNA. The MFE was normalized to remove the bias that a long sequence tends to have lower MFE and hence, the AMFE was calculated by the following equation:

$$AMFE = \frac{-MFE}{L} * 100 \quad (1)$$

where L is the length of the RNA sequence. The MFEI was then calculated using the AMFE by using equation (2):

$$MFEI = \frac{AMFE}{(G+C)\%} \quad (2)$$

RNA motifs

RNA motifs involved in RNA tertiary interactions occur frequently in naturally occurring RNAs.[61] Majority of these structural elements are tetraloops consisting of four nucleotides and generally belong to three classes: UNCG, GNRA and CUUG (where N is any nucleotide and R is a purine). The occurrence frequency of most abundant five motifs (UUCG, GCAA, GAAA, GAGA and CUUG) were compared between different classes of RNAs.

The frequency of each of these five motifs in each sequence of the different datasets of RNAs was calculated. To remove the length bias, the occurrence of a motif was divided by the sequence length of the RNA.

RNA folding measures

The different RNA folding measures such as base-pairing propensity, Shannon entropy and base-pair distance were calculated by the genRNAstats perl script interfaced with the module RNAlib of Vienna RNA Package.[21,60] To remove the bias that a long sequence is likely to have more base pairs, the parameter normalized base-pairing propensity (Nbp) was used. Nbp represents the total number of base pairs present in the RNA secondary structure per unit length of the RNA sequence[62] and ranges from 0 to 0.5; 0 for no base-pair interactions and 0.5 for all base-pairings.

Normalized Shannon entropy (NQ) and normalized base-pair distance (ND) were calculated using the McCaskill base-pair probability (P_{ij}) which measures the probability of base-pairing between bases i and j in a RNA sequence. NQ and ND were calculated using equations 3 and 4 respectively, where L represents the length of the RNA sequence and for values of i and j lies between 1 and L ($1 \le i < j \le L$).

$$NQ = -\frac{1}{L}\sum_{i<j} P_{ij} \ \ log_2\ (P_{ij}) \quad (3)$$

$$ND = \frac{1}{L}\sum_{i<j} P_{ij}\ \left(1 - P_{ij}\right) \quad (4)$$

Random forest classifier

WEKA[32] Version 3.4.3 was used to build a classifier that can differentiate different classes of RNAs. For this, we took a total of 3498 instances (816 sRNAs, 271 rRNAs, 1421 tRNA and 990 mRNA sequences). To overcome the overfitting problem generated due to the imbalance of data in our study (rRNA dataset is smaller as compared to the

other two), we used SMOTE algorithm (Synthetic minority oversampling technique).[63] SMOTE uses an over-sampling approach, where the minority class is over-sampled by creating "synthetic" examples. Application of SMOTE on our dataset increased the rRNA instances from 271 to 542 and the final dataset now comprised of 3769 instances (816 sRNAs, 542 rRNAs, 1421 tRNA and 990 mRNA sequences). The "*Randomize*" filter in WEKA was further applied to shuffle the instances that were created by SMOTE.

We then divided the 3769 instances into two parts: one as *training* dataset, which was used to develop a classifier and the other as *blind test* dataset that was later used to re-evaluate the classifier. 80% (3015 instances) of these instances were randomly chosen for the training dataset while the rest 20% (754 instances) were kept as the blind test dataset. The CfsSubsetEval[64] in the WEKA program was used as an attribute evaluator which listed G+C, Nbp, NQ, AMFE and all motif (UUCG, GCAA, GAAA, GAGA and CUUG) frequency to be the best features to differentiate different classes of RNAs. We thus excluded ND and MFEI, and based on the rest 9 features, a 10-fold cross validation on the training dataset was carried out. In 10-fold cross validation, the training dataset was first randomly divided into 10 equal subsamples, of which 9 subsamples were used as the training set to develop a model, while the one leftout was used to test the model. This process was repeated 10 times with each of the ten subsamples used exactly once in the validation process. Using this technique, a Random Forest (RF) model was developed for classifying different classes of RNAs. The RF model was then re-evaluated on the blind test dataset consisting of 754 instances.

Conflict of interest

No potential conflicts of interest were disclosed.

Funding

This project was supported by Indian Council of Medical Research [extramural project (IRIS ID: 2013–1551G)].

References

1. Waters LS, Storz G. Regulatory RNAs in bacteria. Cell. 2009;136:615–28. doi:10.1016/j.cell.2009.01.043.
2. Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria. Biol Chem. 2005;386:1219–38. doi:10.1515/BC.2005.140.
3. Eddy SR. Noncoding RNA genes. Curr Opin Genet Dev. 1999;9:695–9. doi:10.1016/S0959-437X(99)00022-2.
4. Ortega AD, Quereda JJ, Pucciarelli MG, Garcia-del Portillo F. Non-coding RNA regulation in pathogenic bacteria located inside eukaryotic cells. Front Cell Infect Microbiol. 2014;4:162. doi:10.3389/fcimb.2014.00162.
5. Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell. 2011;43:880–91. doi:10.1016/j.molcel.2011.08.022.
6. Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. Bioinformatics. 2008;24:2807–13. doi:10.1093/bioinformatics/btn560.
7. Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N, Possedko M, Chevalier C, Helfer AC, Benito Y, Jacquier A, et al. Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. Genes Dev. 2007;21:1353–66. doi:10.1101/gad.423507.
8. Wadler CS, Vanderpool CK. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc Natl Acad Sci U S A. 2007;104:20454–9. doi:10.1073/pnas.0708102104.
9. Harris JF, Micheva-Viteva S, Li N, Hong-Geller E. Small RNA-mediated regulation of host-pathogen interactions. Virulence. 2013;4:785–95. doi:10.4161/viru.26119.
10. Li L, Huang D, Cheung MK, Nong W, Huang Q, Kwan HS. BSRD: a repository for bacterial small regulatory RNA. Nucleic Acids Res. 2013;41:D233–8. doi:10.1093/nar/gks1264.
11. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet. 2010;44:167–88. doi:10.1146/annurev-genet-102209-163523.
12. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol. 2011;9:578–89. doi:10.1038/nrmicro2615.
13. Murina VN, Nikulin AD. Bacterial Small Regulatory RNAs and Hfq Protein. Biochem Biokhimiia. 2015;80:1647–54. doi:10.1134/S0006297915130027.
14. Updegrove TB, Zhang A, Storz G. Hfq: the flexible RNA matchmaker. Curr Opin Microbiol. 2016;30:133–8. doi:10.1016/j.mib.2016.02.003.
15. Wassarman KM. Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. Cell. 2002;109:141–4. doi:10.1016/S0092-8674(02)00717-1.
16. Geissmann TA, Touati D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. EMBO J. 2004;23:396–405. doi:10.1038/sj.emboj.7600058.
17. Zhang A, Wassarman KM, Rosenow C, Tjaden BC, Storz G, Gottesman S. Global analysis of small RNA and mRNA targets of Hfq. Mol Microbiol. 2003;50:1111–24. doi:10.1046/j.1365-2958.2003.03734.x.
18. Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet. 2008;4:e1000163. doi:10.1371/journal.pgen.1000163.
19. Pitman S, Cho KH. The Mechanisms of Virulence Regulation by Small Noncoding RNAs in Low GC Gram-Positive Pathogens. Int J Mol Sci. 2015;16:29797–814. doi:10.3390/ijms161226194.
20. Miller EW, Cao TN, Pflughoeft KJ, Sumby P. RNA-mediated regulation in Gram-positive pathogens: an overview punctuated with examples from the group A Streptococcus. Mol Microbiol. 2014;94:9–20. doi:10.1111/mmi.12742.
21. Ng Kwang Loong S, Mishra SK. Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. RNA. 2007;13:170–87. doi:10.1261/rna.223807.
22. 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. doi:10.1016/S0960-9822(01)00401-8.
23. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics. 2001;2:8. doi:10.1186/1471-2105-2-8.
24. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. RNA. 2012;18:900–14. doi:10.1261/rna.029041.111.
25. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. Pac Symp Biocomputing. 2010;15:69–79.
26. 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. doi:10.1093/nar/29.19.3928.
27. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2001;2:919–29. doi:10.1038/35103511.
28. Guo X, Gao L, Wang Y, Chiu DK, Wang T, Deng Y. Advances in long noncoding RNAs: identification, structure prediction and function annotation. Brief Funct Genomics. 2016;15:38–46.
29. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A. 2005;102:2454–9. doi:10.1073/pnas.0409169102.
30. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S. Novel small RNA-encoding genes in the intergenic

regions of Escherichia coli. Curr Biol. 2001;11:941–50. doi:10.1016/S0960-9822(01)00270-6.
31. Nithin C, Patwa N, Thomas A, Bahadur RP, Basak J. Computational prediction of miRNAs and their targets in Phaseolus vulgaris using simple sequence repeat signatures. BMC Plant Biol. 2015;15:140. doi:10.1186/s12870-015-0516-3.
32. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11:10–8. doi:10.1145/1656274.1656278.
33. Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA. Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci. 2006;63:246–54. doi:10.1007/s00018-005-5467-7.
34. Vazquez-Anderson J, Contreras LM. Regulatory RNAs: charming gene management styles for synthetic biology applications. RNA Biol. 2013;10:1778–97. doi:10.4161/rna.27102.
35. Pichon C, Felden B. Intergenic sequence inspector: searching and identifying bacterial RNAs. Bioinformatics. 2003;19:1707–9. doi:10.1093/bioinformatics/btg235.
36. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics. 2000;16:583–605. doi:10.1093/bioinformatics/16.7.583.
37. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci Royal Soc. 2001;268:493–7. doi:10.1098/rspb.2000.1397.
38. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 2004;573:73–7. doi:10.1016/j.febslet.2004.07.056.
39. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997;44:632–6. doi:10.1007/PL00006186.
40. Brennan RG, Link TM. Hfq structure, function and ligand binding. Curr Opin Microbiol. 2007;10:125–33. doi:10.1016/j.mib.2007.03.015.
41. Valentin-Hansen P, Eriksen M, Udesen C. The bacterial Sm-like protein Hfq: a key player in RNA transactions. Mol Microbiol. 2004;51:1525–33. doi:10.1111/j.1365-2958.2003.03935.x.
42. Nielsen JS, Lei LK, Ebersbach T, Olsen AS, Klitgaard JK, Valentin-Hansen P, Kallipolitis BH. Defining a role for Hfq in Gram-positive bacteria: evidence for Hfq-dependent antisense regulation in Listeria monocytogenes. Nucleic Acids Res. 2010;38:907–19. doi:10.1093/nar/gkp1081.
43. Romby P, Charpentier E. An overview of RNAs with regulatory functions in gram-positive bacteria. Cell Mol Life Sci. 2010;67:217–37. doi:10.1007/s00018-009-0162-8.
44. Brantl S, Bruckner R. Small regulatory RNAs from low-GC Gram-positive bacteria. RNA Biol. 2014;11:443–56. doi:10.4161/rna.28036.
45. Jousselin A, Metzinger L, Felden B. On the facultative requirement of the bacterial RNA chaperone, Hfq. Trends Microbiol. 2009;17:399–405. doi:10.1016/j.tim.2009.06.003.
46. Jucker FM, Heus HA, Yip PF, Moors EH, Pardi A. A network of heterogeneous hydrogen bonds in GNRA tetraloops. J Mol Biol. 1996;264:968–80. doi:10.1006/jmbi.1996.0690.
47. Woese CR, Winker S, Gutell RR. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". Proc Natl Acad Sci U S A. 1990;87:8467–71. doi:10.1073/pnas.87.21.8467.
48. Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. The crystal structure of UUCG tetraloop. J Mol Biol. 2000;304:35–42. doi:10.1006/jmbi.2000.4204.
49. Antao VP, Lai SY, Tinoco I, Jr. A thermodynamic study of unusually stable RNA and DNA hairpins. Nucleic Acids Res. 1991;19:5901–5. doi:10.1093/nar/19.21.5901.
50. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10:1178–90. doi:10.1261/rna.7650904.
51. Freyhult E, Gardner PP, Moulton V. A comparison of RNA folding measures. BMC Bioinformatics. 2005;6:241. doi:10.1186/1471-2105-6-241.
52. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6. doi:10.1093/nar/gks1219.
53. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:W686–W9. doi:10.1093/nar/gki366.
54. Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2017;45:D12–D7. doi:10.1093/nar/gkw1071.
55. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2. doi:10.1093/bioinformatics/btq003.
56. Ferrara S, Brugnoli M, De Bonis A, Righetti F, Delvillani F, Deho G, Horner D, Briani F, Bertoni G. Comparative profiling of Pseudomonas aeruginosa strains reveals differential expression of novel unique and conserved small RNAs. PloS One. 2012;7:e36553. doi:10.1371/journal.pone.0036553.
57. 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. doi:10.1073/pnas.112063799.
58. Schattner P. Searching for RNA genes using base-composition statistics. Nucleic Acids Res. 2002;30:2076–82. doi:10.1093/nar/30.9.2076.
59. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26. doi:10.1186/1748-7188-6-26.
60. Hofacker IL. Vienna RNA secondary structure server. Nucleic Acids Res. 2003;31:3429–31. doi:10.1093/nar/gkg599.
61. Moore PB. Structural motifs in RNA. Annual Rev Biochem. 1999;68:287–300. doi:10.1146/annurev.biochem.68.1.287.
62. Schultes EA, Hraber PT, LaBean TH. Estimating the contributions of selection and self-organization in RNA secondary structure. J Mol Evol. 1999;49:76–83. doi:10.1007/PL00006536.
63. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57.
64. Hall MA, Smith LA. Practical feature subset selection for machine learning. Australian Computer Science Conference: Springer, 1998;181–91.