BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2875–2882
doi:10.1093/bioinformatics/btu398

Sequence analysis

Advance Access publication June 23, 2014

A multiobjective method for robust identification of bacterial small non-coding RNAs

Javier Arnedo[1], Rocío Romero-Zaliz[1,2], Igor Zwir[1,2,3] and Coral del Val[1,2,*]

[1]Department of Computer Science and Artificial Intelligence, Universidad de Granada, Granada 18071, Spain, [2]Instituto de Investigación Biosanitaria ibs.GRANADA. Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain and [3]Department of Psychiatry at Washington University, St. Louis, MO 63130, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Small non-coding RNAs (sRNAs) have major roles in the post-transcriptional regulation in prokaryotes. The experimental validation of a relatively small number of sRNAs in few species requires developing computational algorithms capable of robustly encoding the available knowledge and using this knowledge to predict sRNAs within and across species.

Results: We present a novel methodology designed to identify bacterial sRNAs by incorporating the knowledge encoded by different sRNA prediction methods and optimally aggregating them as potential predictors. Because some of these methods emphasize specificity, whereas others emphasize sensitivity while detecting sRNAs, their optimal aggregation constitutes trade-off solutions between these two contradictory objectives that enhance their individual merits. Many non-redundant optimal aggregations uncovered by using multiobjective optimization techniques are then combined into a multiclassifier, which ensures robustness during detection and prediction even in genomes with distinct nucleotide composition. By training with sRNAs in *Salmonella enterica* Typhimurium, we were able to successfully predict sRNAs in *Sinorhizobium meliloti*, as well as in multiple and poorly annotated species. The proposed methodology, like a meta-analysis approach, may begin to lay a possible foundation for developing robust predictive methods across a wide spectrum of genomic variability.

Availability and implementation: Scripts created for the experimentation are available at http://m4m.ugr.es/SupInfo/sRNAOS/sRNAOSscripts.zip.

Contact: delval@decsai.ugr.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2014; revised on June 2, 2014; accepted on June 16, 2014

1 INTRODUCTION

Small non-coding RNAs (sRNAs) have major roles in the bacterial post-transcriptional regulation, affecting important cellular processes such as cell division (Wassarman and Storz, 2000) and response to environmental stimuli (Majdalani and Gottesman, 2005), among others. Experimental methods, including next-generation sequencing technologies, are critical for the functional characterization of sRNAs (Sittka *et al.*, 2009). However, the experimental verification of a relatively small number of sRNAs, primarily in *Escherichia coli* K12 (Argaman *et al.*, 2001; Rivas, 2005) and *Salmonella enterica serovar* Typhimurium LT2 (SLT2) (Padalon-Brauch *et al.*, 2008; Pfeiffer *et al.*, 2007; Sittka *et al.*, 2009; Vogel, 2009), is not sufficient to help detect them in the large list of available sequenced genomes. Therefore, there is a requirement of developing computational algorithms capable of robustly encoding the knowledge of known sRNAs available in certain genomes and using this knowledge to predict sRNAs in other species (Sridhar and Gunasekaran, 2013).

The computational identification of non-coding RNAs (ncRNAs), which includes sRNAs, (ribosomal RNAs) rRNAs and transfer RNAs (tRNAs), is constrained by their diversity in terms of structures, sequences and functions (Livny *et al.*, 2005; Lu *et al.*, 2011; Pichon and Felden, 2003; Rivas and Eddy, 2000, 2001; Storz *et al.*, 2011). Structures of ncRNAs are often scored based on thermodynamic stability, conservation and/or covariance of sequence alignments. Sequences of ncRNAs are primarily analyzed by using two strategies: *ab initio*, which scrutinizes a single-query sequence [e.g. zMFold (Babak *et al.*, 2007; Zuker, 2003), vsFold (Dawson *et al.*, 2007)] and comparative [e.g. QRNA (Rivas and Eddy, 2001), Alifoldz (Washietl and Hofacker, 2004), dynalign (Mathews and Turner, 2002), MSARi (Coventry *et al.*, 2004) and RNAz2 (Gruber *et al.*, 2010)], where the query sequence is investigated by its similarity to other aligned sequences. The latter strategy requires methods for accessing databases, as well as heuristics to efficiently evaluate similarity among sequences. Both strategies require classifiers [e.g. support vector machines (Gruber *et al.*, 2010) or customized hidden Markov models (Rivas and Eddy, 2001)] to predict new ncRNAs based on the acquired knowledge in the training phase. Remarkably, most of the training sets currently used to predict sRNAs are composed of positive examples of rRNAs and tRNAs, which are easier to identify than sRNAs because of their longer sequences and well-defined structure.

Although several methods have been developed and represent a step forward in the computational detection of sRNAs, their success is limited because of (i) the excessive emphasis on specificity, which generates a high number of false-negative predictions by targeting few well-known sRNA families corresponding only to a small percentage of the available sRNAs, (ii) the use of genomic features (e.g. motifs of terminators, RNA polymerase) that are only conserved in closely related organisms [e.g. *E.coli* and *Salmonella* (Argaman *et al.*, 2001; Vogel, 2009)] and thus cause

*To whom correspondence should be addressed.

a large number of false-negative predictions in distantly related organisms and (iii) the emphasis on sensitivity to detect novel findings by thermodynamical approaches, where several examples shown in bacterial genomes harbor well-defined secondary structures and high thermodynamical stability (Eddy, 2001) but do not correspond to true sRNA sequences (Xu *et al.*, 2009) (i.e. predicted false-positive findings). Combining pairs of these methods partially overcome these limitations (del Val *et al.*, 2007; Lu *et al.*, 2011). For example, we previously combined QRNA and RNAz to identify sRNAs in *Sinorhizobium meliloti*; however, the significant reduction in the number of false-positive findings leads to very low sensitivity (del Val *et al.*, 2007).

The performance of a classifier or prediction method may be improved when combining different methods and thus suggests a path to improve bacterial sRNA prediction through meta-analyses of results from existing tools. Therefore, here we propose a new methodology (Fig. 1), which performs optimal aggregations of existing methods—termed basic methods—to predict sRNAs by using machine learning and optimization techniques. This strategy minimizes the false-positive rate during detection and prediction and, simultaneously, maximizes the number of sRNA identified, independently of the evaluated genome. We trained our methodology using the SLT2 genome, which is the *Gammaproteobacteria* model organism harboring the highest number of experimentally validated sRNAs. The resulting strategy was later successfully applied to predict sRNAs in *S.meliloti*, a distant *Alphaproteobacteria* of a great agricultural importance, that has not been used in the training set and has numerous annotated sRNAs (Ulvé *et al.*, 2007; Venkova-Canova *et al.*, 2004). Finally, the strategy was tested in a multispecies dataset (Lu *et al.*, 2011). The performance achieved by our methodology when compared with that of the basic or pairwise

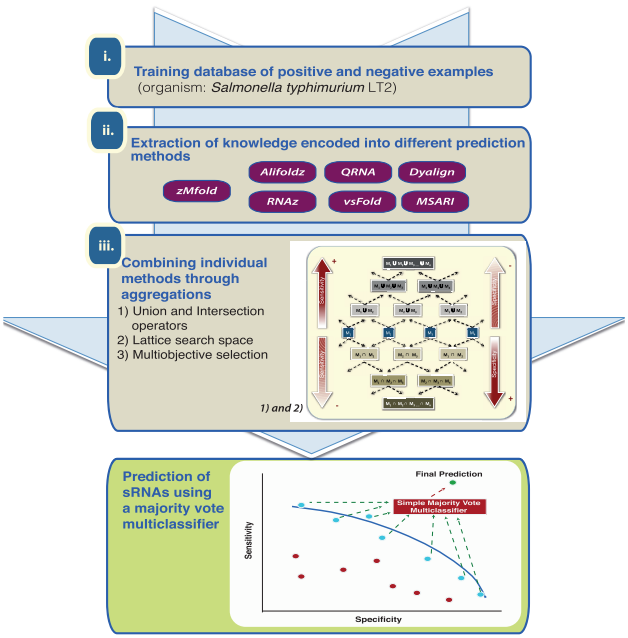

Fig. 1. Prediction of sRNAs by using machine learning and optimization techniques through the optimal aggregation of existing methods termed basic methods

combinations of methods demonstrated that the proposed methodology is accurate and robust to detect sRNAs even in distantly related species. Note that our approach does not invalidate methods developed *de novo* but complements them by providing an efficient way of combining their most reliable features and thus extracts the maximum benefits from each method. Our approach might bring some light into the development of robust predictive methods across a wide spectrum of genomic variability.

2 METHODS

2.1 Selection of basic methods for predicting sRNAs

We selected different methods developed to identify sRNAs as inputs. These include zMFold (Babak *et al.*, 2007; Zuker, 2003), vsFold (Dawson *et al.*, 2007), QRNA (Rivas and Eddy, 2001), Alifoldz (Washietl and Hofacker, 2004), dynalign (Mathews and Turner, 2002), MSARi (Coventry *et al.*, 2004) and RNAz2 (Gruber *et al.*, 2010) (see Supplementary Material for specific description and parameters). They exhibit different characteristics in terms of their subjacent algorithms (i.e. covering all three main approaches to RNA prediction: thermodynamic stability, conservation and covariance of sequence alignments), implementation strategy and subjacent training sets (Supplementary Table S1).

2.2 Creation of positive and negative training datasets

To train our methodology, we created a dataset with both positive and negative examples of sRNAs in the bacterial genome of SLT2 (Supplementary Fig. S1). This organism was selected as a model organism because it harbors a large number of verified sRNAs. The SLT2 dataset of positive examples includes 193 experimentally verified sRNAs (Supplementary Table S2). A total of 52 sRNAs were identified by RNA-seq experiments, whereas the others were identified by traditional experimental approaches (Padalon-Brauch *et al.*, 2008; Papenfort *et al.*, 2008; Pfeiffer *et al.*, 2007; Sittka *et al.*, 2008, 2009) or registered in specialized databases such as RFAM (RFAM version 10.0; Griffiths-Jones *et al.*, 2005). To simulate real prediction conditions, we added the 120 nt upstream and downstream of each sequence in the dataset.

Some of the basic methods selected for this study (Supplementary Table S1) use a comparative strategy and therefore need similar sequences in other species. Our positive dataset includes pairwise or multiple alignment of sequences. These methods are highly dependent on the similarity between the sequences constituting the input alignments. The most informative alignments are composed of sequences sharing between 60 and 85% of similarity (Rivas, 2005). To achieve these standards, we selected sequences from *E.coli* K12, *Klebsiella pneumonia*, *Xylella fastidiosa* and *Yersinia pestis* KIM (genome sequences and annotations were downloaded from the NCBI ftp server, ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Pairwise alignments were performed by using NCBI-BLAST (Altschul *et al.*, 1990) (http://blast.ncbi.nlm.nih.gov/), with a word size of 8, and default parameters. Alignments with an E-value >0.00001 and a length <50 nt were discarded. Multiple alignments were based on previously calculated pairwise alignments by using T-COFFEE (Notredame *et al.*, 2000). This algorithm uses the pairwise alignments to build up a library of alignment information that guides the progressive multiple alignments. This progressive strategy provides a dramatic improvement in accuracy with a modest sacrifice in speed when compared with other multiple alignment alternatives (Notredame *et al.*, 2000).

To estimate false-positive predictions, we generated a dataset of negative examples by shuffling the individual, pairwise and multiple-aligned sequences in the positive dataset. Individual sequences were shuffled using the program shuffleseq (Rice *et al.*, 2000), which is included in the emboss package. This script takes an input sequence and randomly shuffles the order of its base without affecting the composition. Aligned

sequences were randomly shuffled using the shuffle-pair.pl method (Babak *et al.*, 2007). This method preserves the properties of the original sequences or alignments (e.g. overall guanine-cytosine (GC) content, % identity, distributions of identities, mismatches and gaps), while destroys correlated base-pairing patterns in a conserved RNA structure. An alternative approach is to consider genome regions without annotated sRNAs as negative examples (Lu *et al.*, 2011). However, absence of knowledge does not imply knowledge of absence, and there are many examples in our previous work in transcriptional regulation where non-meaningful genomic regions resulted in truly functional genomic regions (Zwir *et al.*, 2005). Therefore, we believe that a synthetic negative dataset based on shuffling of the positive dataset sequences secures the disruption of secondary structures implicated in sRNAs and thus may produce less uncertainty about negative examples and better estimation of specificity. However, the use of non-annotated sequences as negative dataset would be a better choice if a more conservative estimate of the precision is envisaged.

2.3 Aggregations of basic methods

The aggregation of methods is performed by systematically applying the union –∪– and intersection –∩– operations (Supplementary Fig. S2) to sets of predictions performed by basic or previous aggregations of methods (Fig. 1). These aggregations are reminiscent of logic expressions representing combined predictions. The union operation ($method_A \cup method_B$) reports the largest subsequences predicted or covered by any of the operands (Supplementary Fig. S3a). The intersection operation ($method_A \cap method_B$) reports a sequence embracing only those nucleotides predicted by both operands (Supplementary Fig. S3b).

2.4 Evaluation of methods: sensitivity and specificity

The evaluation of results from basic or aggregation of methods was carried out in a similar fashion. A query sequence in either the positive or the negative dataset is considered to be predicted by a method if (i) the predicted sequence is included in the query sequence, (ii) the query sequence is included in the predicted sequence or (iii) at least 70% of the nucleotides in the query sequence are covered by the predicted sequence. True-positive predictions (TP) are defined as the number of sequences in the positive dataset predicted by a method; true-negative predictions (TN) correspond to the number of sequences in the negative dataset that are not predicted by a method; false-positive predictions (FP) correspond to the number of sequences in the negative dataset that are predicted by a method; and false-negative predictions (FN) is the number of sequences in the positive dataset that are not predicted by a method. Sensitivity is defined as the proportion of sequences in the positive dataset that are predicted by a method [Sn; Equation (1)], whereas specificity [Sp; Equation (2)] corresponds to the proportion of sequences in the negative dataset that were not predicted by a method.

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TN}{TN + FP} \quad (2)$$

2.5 Identification of optimal aggregation of methods

To optimally aggregate methods and their corresponding predictions of sRNAs, we implemented an efficient multiobjective evolutionary algorithm relying on the NSGA-II algorithm (see Supplementary Material). Optimal aggregations are defined as those that maximize two objectives: Sp and Sn. To prevent data overfitting and bloat (Deb *et al.*, 2002), we applied the following constrains to select specific aggregations of methods: (i) if two aggregation of methods have the same Sp and Sn, the one with the lower number of basic methods is preferred, and (ii) a basic method might only appear one time in each aggregation of methods.

2.6 Combination of aggregations of methods in a simple majority voting multiclassifier

We combined the optimal aggregations of methods by using a majority voting strategy (Supplementary Fig. S4). This strategy was selected because of its simplicity, high level of accuracy and robustness (Belaid and Anigbogu, 1994; Lam and Suen, 1997; Rahman *et al.*, 2002). The voting procedure is implemented as follows: given n independent experts having the same probability of being correct (in our case, aggregations of methods), and that each of these experts produces a decision regarding the identity of an unknown observation, then the observation is assigned to the class that achieves the highest levels of consensus. To avoid possible sampling biases, we performed 10-fold cross-validation in the training set composed of 193 SLT2 positive sequences. This positive dataset was randomly partitioned into 10 subsets of 20 sequences each. In each round of cross-validation, one set was retained as the validation data for testing the model, and 173 sequences in the remaining 9 subsets are used as training data. The cross-validation process was then repeated 10 times (Supplementary Fig. S5). The Sp and Sn were averaged over the 10-fold cross-validations and reported. The same procedure was carried out for the negative dataset.

2.7 Testing the multiclassifier performance by predicting sRNAs in *S.meliloti* and in a multispecies dataset

To test our method with sRNAs that have not been included in our original training set, we created a new dataset of experimentally validated sRNAs similar to that developed for SLT2 but from the phylogenetically distant *Alphaproteobacteria S.meliloti* (Supplementary Table S3). This dataset is composed of 81 sRNAs obtained by deep-sequencing techniques (Schlüuter *et al.*, 2010) and other individually validated sequences (del Val *et al.*, 2007; Ulvé *et al.*, 2007; Venkova-Canova *et al.*, 2004). Pairwise and multiple alignments were obtained analogously to those from SLT2 using BLAST and T-COFFEE (Altschul *et al.*, 1990; Notredame *et al.*, 2000), respectively. The performance of our method was also evaluated with a wide variety of organisms by using a multispecies dataset (Lu *et al.*, 2011), which was previously used in predicting sRNAs. This dataset is composed of 776 putative sRNAs from 14 different genomes including *Helicobacter pylori*, *Staphylococcus aureus*, *Bacillus subtilis*, *Listeria monocytogenes*, *Chlamydia trachomatis*, *Shewanella oneidensis*, *Xenorhabdus nematophila*, *Vibrio cholerae*, *E.coli*, SLT2, *Pseudomonas aeruginosa*, *Caulobacter crescentus*, *Burkholderia cenocepacia* and *Streptomyces coelicolor A3*. The sRNAs in this dataset were obtained by experimental validation (Huang *et al.*, 2009; Sittka *et al.*, 2008), genome-tiling microarray experiments (Toledo-Arana *et al.*, 2009) and RNA-seq experiments (Albrecht *et al.*, 2010; Liu *et al.*, 2009; Lu *et al.*, 2011; Sharma *et al.*, 2010; Sittka *et al.*, 2008; Yoder-Himes *et al.*, 2009).

3 RESULTS

We applied the basic methods (Supplementary Table S1) and the proposed methodology (Fig. 1) to identify sRNAs in the positive and negative SLT2 datasets of sRNAs (see Section 2). Then, we predicted sRNAs in a new dataset of experimentally validated sRNAs (del Val *et al.*, 2007; Venkova-Canova *et al.*, 2004; Ulvé *et al.*, 2007) from *S.meliloti*, as well as in a multispecies dataset (Lu *et al.*, 2011) (see Section 2).

3.1 Basic methods identify sRNAs with disparate sensitivity and specificity

We applied seven basic methods including zMFold, vsFold, QRNA, Alifoldz, dynalign, MSARi and RNAz2 (Supplementary Table S1) to identify sRNAs. These methods constitute the state-of-the-art in the field and implement diverse methodologies that embrace the most relevant features that characterize sRNAs. The performance of the basic methods was estimated according to their Sp and Sn by scoring the sequences/alignments from the positive and negative SLT2 training datasets. The obtained results reveal that these methods tend to favor either Sp or Sn, but not both of them simultaneously (Table 1). For example, RNAz2 (Sp 0.98, Sn 0.27) exhibits high Sp but low Sn, whereas zMFold (Sp 0.49, Sn 0.90) does the opposite. QRNA achieves the best trade-off between both Sp and Sn. We found that most of the sRNAs are predicted by more than two basic methods (Supplementary Fig. S6). A large group of sRNAs is predicted by just one method (55 sequences) but with low Sp, whereas only a small group of sequences is not predicted by any of the methods (8 sequences; Supplementary Fig. S6). These differences in terms of Sp and Sn exhibited by the basic methods—trained mainly with *S.enterica* Thyphimurium and *E.coli*—raise the question if their combination could be successfully applicable to other genomes.

3.2 The proposed strategy identifies optimal aggregations of basic methods that predict sRNAs

The basic methods showed different and sometimes opposing Sp and Sn values when predicting sRNAs (Table 1). This may be due to their specific methodological approaches that tend to emphasize particular features that characterize sRNAs but not others. To address this problem, we incorporated and combined the particular skills of each basic predictor and eliminated their contradictory information. To combine the most promising characteristics of basic methods, we run all basic methods over the training dataset (i.e. SLT2 dataset) and searched for optimal aggregations of these methods. The aggregations are performed by combining basic methods using the union –∪– and the intersection –∩– operations (Halmos, 1961) on their predictions at the nucleotide level (Supplementary Fig. S3). These typical mathematical set operations have their parallel logic operations (i.e. OR and AND), which consecutive application allows defining any possible logic expression. The potential aggregation of methods (here $>20\,000$) creates a large space of potential hypotheses, which can be represented as a lattice (Supplementary Fig. S2). Each aggregation in the lattice can be considered as a new prediction method.

Table 1. Sensitivity and specificity of the individual methods for the SLT2 dataset

Description	SLT2 specificity	SLT2 sensitivity
RNAz2	0.98	0.27
vsFold	0.88	0.25
Alifoldz	0.87	0.42
Dynalign	0.86	0.28
QRNA	0.71	0.59
MSARi	1.00	0.02
zMFold	0.49	0.90

The union of methods increases Sn and decreases Sp, whereas the intersection does the opposite, as expected. Then, the combination of union and intersection operations in the same aggregation may increase Sp and Sn simultaneously. However, the aggregation of all methods does not necessarily ensure more accurate predictions (Cordón *et al.*, 2002). For instance, two methods may have contradictory predictive strategies, or even more, the simple summation of them can lead to overfitting (Cordón *et al.*, 2002; Harari *et al.*, 2010; Zwir *et al.*, 2005). Moreover, the computational complexity of performing an exhaustive search of all logical expressions (i.e. Boolean satisfiability problem) is of complexity NP-complete (Gu *et al.*, 1996). Therefore, a careful optimization strategy is required to integrate the basic methods in an appropriate fashion.

To address the computational complexity described above, our methodology applies a heuristic approach consisting in the generation of optimal aggregations by using multiobjective optimization techniques; particularly, we used the genetic algorithm NSGA-II (Supplementary Material) that simultaneously search for trade-off solutions between two objectives (i.e. Sp and Sn). Because there are two possibly contradictory objectives to be optimized simultaneously, there is no single optimal solution but rather a family of optimal results (Table 2). These results are organized as a Pareto optimal front (Supplementary Fig. S5). This front includes all non-dominated solutions (Deb, 2001; Ruspini and Zwir, 2002), where one solution is said to dominate another solution when it is better than the other in all objectives being considered (e.g. both sensitivity and specificity).

3.2.1 Optimal aggregations of basic methods enhance their specificity and sensitivity for identifying sRNAs We identified 26 different optimal aggregations based on the SLT2 dataset that correspond to non-dominated solutions (Table 2 and Supplementary Fig. S7). For example, both aggregations AGR_14 (Sp 0.79, Sn 0.66) and AGR_26 (Sp 0.33, Sn 0.98) are in the Pareto front of optimal solutions because the first has better Sp values but worst Sn values than the second. Of all these aggregations, the ones with the best possible trade-offs among the selected objectives normally lie on the 'knee' of the Pareto front (Deb, 2001) (Supplementary Fig. S4). All 26 aggregations dominate the basic methods based on the Sp and Sn objectives, and thus none of the latter solutions are in the optimal Pareto front (Supplementary Fig. S4). The methods zMfold and RNAz2 are always present in aggregations achieving the best Sn scores ($Sn>0.93$), and their corresponding Sp scores are between 0.33 and 0.49 [e.g. (((zMFold ∪ QRNA) ∪ Alifoldz) ∪ RNAz2); ((RNAz2 ∪ zMFold) ∪ dynalign)]. The aggregations harboring the best Sp scores ($Sp>0.93$) are mainly composed of at least one intersection of basic methods harboring high Sn combined with the union of methods displaying high Sp [e.g. ((QRNA ∩ RNAz2) ∪ MSARi); ((((vsFold ∪ Alifoldz) ∩ (QRNA ∩ zMFold)) ∪ RNAz2) ∪ MSARi)], and their Sn scores stand between 0.25 and 0.49. Finally, there is not obvious pattern in the most balanced aggregations (i.e. Sp and Sn ca. 0.7).

Table 2. Sensitivity and specificity of the non-dominated aggregations of methods obtained and our methodology result for the *Salmonella enterica* serovar Typhimurium LT2 (SLT2) dataset

ID	Description	SLT2 specificity	SLT2 sensitivity
AGR_1	((QRNA ∩ RNAz2) ∪ MSARi)	1.00	0.25
AGR_2	(RNAz2 ∪ MSARi)	0.98	0.28
AGR_3	(RNAz2 ∪ (((QRNA ∩ Alifoldz) ∩ vsFold) ∪ MSARi))	0.97	0.34
AGR_4	(RNAz2 ∪ (((QRNA ∩ Alifoldz) ∩ zMFold) ∪ MSARi))	0.96	0.40
AGR_5	((((QRNA ∩ zMFold) ∩ vsFold) ∪ RNAz2) ∪ MSARi)	0.95	0.42
AGR_6	((((vsFold ∪ Alifoldz) ∩ (QRNA ∩ zMFold)) ∪ RNAz2) ∪ MSARi)	0.93	0.49
AGR_7	((((vsFold ∪ dynalign) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.88	0.51
AGR_8	((((dynalign ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.87	0.53
AGR_9	((((vsFold ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.85	0.59
AGR_10	(((QRNA ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.83	0.60
AGR_11	((RNAz2 ∪ MSARi) ∪ ((zMFold ∪ Alifoldz) ∩ QRNA))	0.81	0.61
AGR_12	((RNAz2 ∪ MSARi) ∪ ((zMFold ∪ dynalign) ∩ QRNA))	0.81	0.61
AGR_13	((RNAz2 ∪ MSARi) ∪ ((QRNA ∪ dynalign) ∩ zMFold))	0.80	0.62
AGR_14	(RNAz2 ∪ ((QRNA ∪ Alifoldz) ∩ zMFold))	0.79	0.66
AGR_15	(((QRNA ∪ ((vsFold ∪ RNAz2) ∪ MSARi)) ∪ Alifoldz) ∩ zMFold)	0.75	0.67
AGR_16	((RNAz2 ∪ vsFold) ∪ ((QRNA ∪ Alifoldz) ∩ zMFold))	0.70	0.72
AGR_17	((RNAz2 ∪ QRNA) ∪ ((dynalign ∪ Alifoldz) ∩ zMFold))	0.63	0.73
AGR_18	((((vsFold ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ QRNA)	0.61	0.75
AGR_19	(((vsFold ∪ QRNA) ∪ Alifoldz) ∪ RNAz2)	0.56	0.76
AGR_20	((((RNAz2 ∪ vsFold) ∪ QRNA) ∪ Alifoldz) ∪ dynalign)	0.51	0.76
AGR_21	((RNAz2 ∪ ((QRNA ∩ vsFold) ∩ Alifoldz)) ∪ zMFold)	0.49	0.93
AGR_22	(((RNAz2 ∪ zMFold) ∪ MSARi) ∪ (((QRNA ∩ Alifoldz) ∪ dynalign) ∩ vsFold))	0.48	0.94
AGR_23	((RNAz2 ∪ zMFold) ∪ dynalign)	0.43	0.95
AGR_24	(((zMFold ∪ (QRNA ∩ dynalign)) ∪ Alifoldz) ∪ RNAz2)	0.42	0.96
AGR_25	((RNAz2 ∪ QRNA) ∪ zMFold)	0.36	0.97
AGR_26	(((zMFold ∪ QRNA) ∪ Alifoldz) ∪ RNAz2)	0.33	0.98
sRNA_OS	Our methodology	0.78	0.67

3.3 Incorporation of all optimal aggregations in a multiclassifier establishes a robust and accurate predictor of sRNAs

Our strategy assembles all optimal aggregations of methods in a multiclassifier, which is a decision-making method that considers all of these solutions in a cooperative strategy toward identifying sRNAs (see Section 2). Here, the cooperation policy was implemented as a simple majority voting approach of all non-dominated aggregations, which guarantees robust predictions (Supplementary Fig. S4). First, we tested the performance of the multiclassifier with respect to the SLT2 training dataset, obtaining 0.78 and 0.67 averaged scores of Sp and Sn, respectively (Table 2). To account for the sample variability, we performed a 10-fold cross-validation on the same SLT2 training dataset (Supplementary Fig. S5 and Supplementary Tables S4–S7), obtaining 0.81 and 0.61 averaged scores of Sp and Sn, respectively. Remarkably, the average of the nine partitions used as training sets along the 10-folds exhibited 0.83 and 0.62 averaged scores of Sp and Sn, respectively. The similar training and test across the cross-validation process exhibits the robustness and stability characteristics of the multiclassifier implemented in our methodology. At least four classes of aggregations identify distinct subsets of sRNAs (Supplementary Fig. S7). One class of aggregations recognizes most of the sRNAs with high Sn but relatively low Sp scores (Table 2; AGR_21 to AGR_26). In contrast, another class of aggregations recognizes only a single subset of sRNAs with high Sp but low Sn (Table 2; AGR_1 to AGR_9). Most of the other aggregations recognize a large subset of sRNAs (Table 2). Notably, the majority voting strategy implemented in our methodology allows preserving the trade-offs between these classes in the prediction process, which is consistent with the criteria used for the selection of aggregations in the Pareto front. Other approaches like predicting by the Maximum Sn or Sp may provide better but less robust results.

Then, we tested the performance of the multiclassifier on another dataset of sRNA examples that were not included in the previous SLT2 training set. As in the case of the SLT2 dataset, we compiled a new dataset 81 experimentally validated sRNAs in the well-studied but phylogenetically distant bacterium *S.meliloti* (see Section 2; Supplementary Table S3). The multiclassifier based on 26 aggregations obtained 0.72 and 0.58 scores of Sp and Sn, respectively (Table 3), whereas the best basic method for this dataset (zMfold) showed 0.44 and 0.67 scores of Sp and Sn, respectively (Table 4). Because the Sp and Sn scores obtained by our methodology are not significantly different from those obtained with the SLT2 training dataset, the obtained results confirm that the proposed strategy is sufficiently robust to make accurate predictions even under different nucleotide composition of the target genome.

We also compared the performance of our methodology with that of the SIPHT method (Livny *et al.*, 2006), which is widely

Table 3. Sensitivity and specificity of the non-dominated aggregations of methods and our methodology result for the *S.meliloti* (SM) dataset

ID	Description	SM specificity	SM sensitivity
AGR_01	((QRNA ∩ RNAz2) ∪ MSARi)	1.00	0.13
AGR_02	(RNAz2 ∪ MSARi)	1.00	0.18
AGR_03	(RNAz2 ∪ (((QRNA ∩ Alifoldz) ∩ zMFold) ∪ MSARi))	1.00	0.28
AGR_04	(RNAz2 ∪ ((Alifoldz ∩ vsFold) ∪ MSARi))	0.89	0.31
AGR_05	((((QRNA ∩ zMFold) ∩ vsFold) ∪ RNAz2) ∪ MSARi)	0.93	0.42
AGR_06	((((vsFold ∪ Alifoldz) ∩ (QRNA ∩ zMFold)) ∪ RNAz2) ∪ MSARi)	0.93	0.44
AGR_07	((((vsFold ∪ dynalign) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.52	0.73
AGR_08	((((dynalign ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.75	0.54
AGR_09	((((vsFold ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.53	0.70
AGR_10	(((QRNA ∩ zMFold) ∪ RNAz2) ∪ MSARi)	0.90	0.48
AGR_11	((RNAz2 ∪ MSARi) ∪ ((zMFold ∪ Alifoldz) ∩ QRNA))	0.88	0.49
AGR_12	((RNAz2 ∪ MSARi) ∪ ((zMFold ∪ dynalign) ∩ QRNA))	0.88	0.57
AGR_13	((RNAz2 ∪ MSARi) ∪ ((QRNA ∪ dynalign) ∩ zMFold))	0.75	0.54
AGR_14	(RNAz2 ∪ ((QRNA ∪ Alifoldz) ∩ zMFold))	0.81	0.52
AGR_15	(((QRNA ∪ ((vsFold ∪ RNAz2) ∪ MSARi)) ∪ Alifoldz) ∩ zMFold)	0.50	0.64
AGR_16	((RNAz2 ∪ vsFold) ∪ ((QRNA ∪ Alifoldz) ∩ zMFold))	0.36	0.94
AGR_17	((RNAz2 ∪ QRNA) ∪ ((dynalign ∪ Alifoldz) ∩ zMFold))	0.67	0.68
AGR_18	((((vsFold ∪ Alifoldz) ∩ zMFold) ∪ RNAz2) ∪ QRNA)	0.46	0.84
AGR_19	(((vsFold ∪ QRNA) ∪ Alifoldz) ∪ RNAz2)	0.27	0.98
AGR_20	((((RNAz2 ∪ vsFold) ∪ QRNA) ∪ Alifoldz) ∪ dynalign)	0.26	0.98
AGR_21	((RNAz2 ∪ ((QRNA ∩ vsFold) ∩ Alifoldz)) ∪ zMFold)	0.44	0.76
AGR_22	(((RNAz2 ∪ zMFold) ∪ MSARi) ∪ (((QRNA ∩ Alifoldz) ∪ dynalign) ∩ vsFold))	0.39	0.84
AGR_23	((RNAz2 ∪ zMFold) ∪ dynalign)	0.36	0.88
AGR_24	(((zMFold ∪ (QRNA ∩ dynalign)) ∪ Alifoldz) ∪ RNAz2)	0.33	0.86
AGR_25	((RNAz2 ∪ QRNA) ∪ zMFold)	0.40	0.86
AGR_26	(((zMFold ∪ QRNA) ∪ Alifoldz) ∪ RNAz2)	0.33	0.88
sRNA_OS	Our methodology	0.72	0.58

used by the scientific community. This method uses the QRNA basic method combined with other computational tools that identify and annotate sequence-based motifs of transcription factor binding sites and rho-independent terminators. Because the SIPHT method uses genomic features for its predictions of sRNAs, our database of negative examples cannot be used for comparison purposes because the motifs of these features are likely to be destroyed in shuffled sequences (see Section 2). Therefore, the Sp score cannot be fairly estimated. Consequently, we compared our methodology with SIPHT in terms of Sn. The predictions of SIPHT for SLT2 and *S.meliloti* were obtained from (http://newbio.cs.wisc.edu/sRNA/). We predict sRNAs with 0.67 and 0.57 averaged scores of Sn in SLT2 and in *S.meliloti*, respectively. In comparison, SIPHT predicts the sRNAs contained in our SLT2 and in *S.meliloti* databases of positive examples with 0.46 and 0.33 scores of Sn, respectively.

We also tested the predictive power of our approach with another dataset composed of sRNAs derived from 14 different genomes (Lu *et al.*, 2011) (see Section 2), which also were not contained in the training set. We significantly overcame the basic methods used by (Lu *et al.*, 2011) in all genomes independently of their GC content, except for *Burkholderia* and *Chlamydia* (Lu *et al.*, 2011). The best performing individual methods in (Lu *et al.*, 2011) obtained 0.27, 0.20, 0.40 and 0.49 of Sn in the different datasets, while our methodology obtained an average of 0.70 Sn across all datasets (Supplementary Table S8). Remarkably, these values were similar in SLT2 and *S.meliloti*.

4 DISCUSSION

The use of next-generation technologies, such as RNA-seq, demonstrated that the number and diversity of sRNAs is greater than was originally expected (Toledo-Arana *et al.*, 2009; Vogel, 2009). Therefore, there is increased interest in identifying sRNAs and deciphering their role in regulatory systems within a particular species or across multiple species. These experimental methods are critical for functional characterization of sRNAs (Sittka *et al.*, 2009); however, their applicability still has been constrained to a relatively small number of sRNAs as well as genomes. Consequently, computational prediction of sRNAs is required to develop new hypothesis that allows focusing the experimental verification on particular targets, to provide clues about the molecular mechanisms governing gene regulation.

Different strategies have been implemented using distinct computational methods to predict sRNAs (Livny *et al.*, 2008; Rivas and Eddy, 2001; Washietl *et al.*, 2005); however, most of them exhibit similar limitations. Essentially, there is a trend in these computational methods to favor either Sp or Sn, but not both, in their predictions (Lu *et al.*, 2011) and that, in turn, generates either a high number of false-positive predictions or false-negative predictions. This is true even if some methods combine in one predictor different sRNA features and/or genomic information (Livny *et al.*, 2005, 2008). In this work, we presented a new methodology, which identifies bacterial sRNAs by simultaneously minimizing the false-positive predictions (Sp) and maximizing the number of recognized sRNAs (Sn). This method is

Table 4. Sensitivity and specificity values of the individual methods for the *S.meliloti* (SM) dataset

Method	SM specificity	SM sensitivity
RNAz2	1.00	0.18
vsFold	0.38	0.84
Alifoldz	0.83	0.31
dynalign	0.70	0.51
QRNA	0.85	0.52
MSARi	1.00	0.00
zMFold	0.44	0.67

able to predict sRNAs in different genomes, even when there is a lack of reliable annotations, because it does not rely on additional genomic features.

Several characteristics distinguish our methodology from other methods. The proposed approach uses the distinctive features of different methods—termed basic methods—instead of developing a new method *de novo*, and combines them in a manner that resolves the problem of contradictory knowledge and thus improving their predictive power. To do that, our methodology combines the predictions of the basic methods by using typical set theory operations such as the union and/or intersection. Likewise in logic expressions, the systematic application of these operations produces chained and disparate aggregations of methods. Because not all aggregations may perform better than an individual method (del Val *et al*., 2007), and large aggregations including all methods may also produce overfitting, our methodology selects optimal aggregations as a trade-off between Sp and Sn by using multiobjective optimization heuristic techniques. Particularly, we used the genetic algorithms NSGA-II. The efficient heuristics used by our methodology avoids intractable processing times, which are common in combinatorial optimization (De Smet and Marchal, 2010). This process results in non-redundant optimal aggregations, where the balance between the two contradictory objectives also prevents overfitting (Cordón *et al*., 2002; Romero-Zaliz *et al*., 2009). This approach substantially improves both Sp and Sn over that of previous single aggregation of two predefined methods (Livny *et al*., 2008; Pichon and Felden, 2003).

We applied our methodology to a dataset of experimentally validated sRNAs in SLT2. The results are significantly better than those results from the basic methods alone (Tables 1 and 2) in training or test datasets (Supplementary Fig. S3s and Supplementary Tables S4–S7). Remarkably, the Sp and Sn scores are similar between both training and test sets, and even within the training test fold partitions performed in the cross-validation. These results strongly suggest that the method is robust for predicting sRNAs despite the possible sample variability. To effectively confirm these findings, we tested the predictive power of the method in two datasets that were not used in the training process: the *S.meliloti* and the multispecies datasets. Despite the different nucleotide composition of these genomes (see GC content; Supplementary Table S8), our methodology obtained good Sp and Sn scores. Finally, our proposed method obtained better Sn scores than SIPHT (Livny *et al*., 2008), a widely used method that uses specific genomic information about binding sites and terminators (Li *et al*., 2012). Unlike our method that uses general characteristics of sRNAs, SIPHT uses particular genomic features. Therefore, the database of negative examples based on shuffled sequences cannot be fairly used to estimate comparable Sps for both methods. It would be interesting to derive negative examples acceptable to both methods to better compare their power.

In sum, our approach has demonstrated to encode a successful and robust methodology to predict sRNAs even in poorly annotated genomes. Moreover, we have shown that appropriately combining results from existing methods in a meta-analysis-like approach may significantly improve their accuracy and facilitates the generation of new predictors by simply using different training datasets and/or aggregating new basic methods.

ACKNOWLEDGEMENT

Author would like to thank Henry V. Huang for the helpful comments on this article.

Funding: All authors were funded in part by FEDER fonds; the Spanish Ministry of Science and Technology under projects TIN2009-13950 and TIN2012-38805; the Consejería de Innovación, Investigación y Ciencia, Junta de Andalucía, under project TIC-02788. C. del Val was financed by the "Plan Propio de Investigación 2013" of the University of Granada.

Conflict of Interest: none declared.

REFERENCES

Albrecht,M. *et al.* (2010) Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res.*, **38**, 868–877.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Argaman,L. *et al.* (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.

Babak,T. *et al.* (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC bioinformatics*, **8**, 33.

Belaid,A. and Anigbogu,J.C. (1994) Use of many classifiers for multifont text recognition. *Trait. Signal*, **11**, 57.

Cordón,O. *et al.* (2002) Linguistic modeling by hierarchical systems of linguistic rules. *IEEE Trans. Fuzzy Syst.*, **10**, 2–20.

Coventry,A. *et al.* (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 12102–12107.

Dawson,W.K. *et al.* (2007) Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. *PloS One*, **2**, e905.

De Smet,R. and Marchal,K. (2010) Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, **8**, 717–729.

Deb,K. (2001) *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester.

Deb,K. *et al.* (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, **6**, 182–197.

del Val,C. *et al.* (2007) Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium* meliloti by comparative genomics. *Mol. Microbiol.*, **66**, 1080–1091.

Eddy,S. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.

Griffiths-Jones,S. *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.

Gruber,A.R. *et al.* (2010) Rnaz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, **2010**, 69–79.

Gu,J. *et al.* (1996) Algorithms for the satisfiability (sat) problem: a survey. In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, pp. 19–152.

Halmos,P.R. (1961) Naive Set Theory. *Proc. Edinb. Math. Soc.*, **12**, 159.

Harari,O. *et al.* (2010) Defining the plasticity of transcription factor binding sites by Deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. *PLoS Comput. Biol.*, **6**, e1000862.

Huang,H.-Y. *et al.* (2009) sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res.*, **37**, D150–D154.

Lam,L. and Suen,S.Y. (1997) Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.*, **27**, 553–568.

Li,W. *et al.* (2012) Predicting sRNAs and their targets in bacteria. *Genomics Proteomics Bioinformatics*, **10**, 276–284.

Liu,J.M. *et al.* (2009) Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res.*, **37**, e46.

Livny,J. *et al.* (2005) sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res.*, **33**, 4096–4105.

Livny,J. *et al.* (2006) Identification of 17 Pseudomonas aeruginosa sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res.*, **34**, 3484–3493.

Livny,J. *et al.* (2008) High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PloS One*, **3**, e3197.

Lu,X. *et al.* (2011) Assessing computational tools for the discovery of small RNA genes in bacteria. *RNA*, **17**, 1635–1647.

Majdalani,N. and Gottesman,S. (2005) The Rcs phosphorelay: a complex signal transduction system. *Ann. Rev. Microbiol.*, **59**, 379–405.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Padalon-Brauch,G. *et al.* (2008) Small RNAs encoded within genetic islands of *Salmonella typhimurium* show host-induced expression and role in virulence. *Nucleic Acids Res.*, **36**, 1913–1927.

Papenfort,K. *et al.* (2008) Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol. Microbiol.*, **68**, 890–906.

Pfeiffer,V. *et al.* (2007) A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the *Salmonella* core genome. *Mol. Microbiol.*, **66**, 1174–1191.

Pichon,C. and Felden,B. (2003) Intergenic sequence inspector: searching and identifying bacterial RNAs. *Bioinformatics*, **19**, 1707–1709.

Rahman,A. *et al.* (2002) Multiple classifier combination for character recognition: revisiting the majority voting system and its variations. In: *Proceedings of the 5th International Workshop on Document Analysis*. pp. 167–178.

Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 2–3.

Rivas,E. (2005) Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics*, **6**, 63.

Rivas,E. and Eddy,S. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Romero-Zaliz,R. *et al.* (2009) Optimization of multi-classifiers for computational biology: application to gene finding and expression. *Theor. Chem. Acc.*, **125**, 599–611.

Ruspini,E. and Zwir,I. (2002) Automated generation of qualitative representations of complex objects by hybrid soft-computing methods. In: *Pattern recognition: from classical to modern approaches*. World Scientific, New Jersey.

Schlüuter,J.P. *et al.* (2010) A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium Sinorhizobium meliloti. *BMC Genomics*, **11**, 45.

Sharma,C.M. *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, **464**, 250–255.

Sittka,A. *et al.* (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.*, **4**, e1000163.

Sittka,A. *et al.* (2009) Deep sequencing of *Salmonella* RNA associated with heterologous Hfq proteins *in vivo* reveals small RNAs as a major target class and identifies RNA processing phenotypes. *RNA Biol.*, **6**, 266–275.

Sridhar,J. and Gunasekaran,P. (2013) Computational small RNA prediction in bacteria. *Bioinform. Biol. Insights*, **7**, 83.

Storz,G. *et al.* (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, **43**, 880–891.

Toledo-Arana,A. *et al.* (2009) The Listeria transcriptional landscape from saprophytism to virulence. *Nature*, **459**, 950–956.

Ulvé,V.M. *et al.* (2007) Identification of chromosomal alpha-proteobacterial small RNAs by comparative genome analysis and detection in *Sinorhizobium* meliloti strain 1021. *BMC Genomics*, **8**, 467.

Venkova-Canova,T. *et al.* (2004) Two discrete elements are required for the replication of a repABC plasmid: an antisense RNA and a stem-loop structure. *Mol. Microbiol.*, **54**, 1431–144.

Vogel,J. (2009) A rough guide to the non-coding RNA world of *Salmonella*. *Mol. Microbiol.*, **71**, 1–11.

Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–30.

Washietl,S. *et al.* (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Wassarman,K.M. and Storz,G. (2000) 6S RNA regulates *E. coli RNA polymerase activity*. *Cell*, **101**, 613–623.

Xu,X. *et al.* (2009) Discovering cis-regulatory RNAs in *shewanella* genomes by support vector machines. *PLoS Comput. Biol.*, **5**, e1000338.

Yoder-Himes,D.R. *et al.* (2009) Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 3976–3981.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

Zwir,I. *et al.* (2005) Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation. *Bioinformatics*, **21**, 4073–4083.