3928–3938 *Nucleic Acids Research, 2001, Vol. 29, No. 19*

A computational approach to identify genes for functional RNAs in genomic sequences

Richard J. Carter, Inna Dubchak[1] and Stephen R. Holbrook*

Computational and Theoretical Biology Department, Physical Biosciences Division and [1]National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Received July 2, 2001; Revised and Accepted August 14, 2001

ABSTRACT

Currently there is no successful computational approach for identification of genes encoding novel functional RNAs (fRNAs) in genomic sequences. We have developed a machine learning approach using neural networks and support vector machines to extract common features among known RNAs for prediction of new RNA genes in the unannotated regions of prokaryotic and archaeal genomes. The *Escherichia coli* genome was used for development, but we have applied this method to several other bacterial and archaeal genomes. Networks based on nucleotide composition were 80–90% accurate in jackknife testing experiments for bacteria and 90–99% for hyperthermophilic archaea. We also achieved a significant improvement in accuracy by combining these predictions with those obtained using a second set of parameters consisting of known RNA sequence motifs and the calculated free energy of folding. Several known fRNAs not included in the training datasets were identified as well as several hundred predicted novel RNAs. These studies indicate that there are many unidentified RNAs in simple genomes that can be predicted computationally as a precursor to experimental study. Public access to our RNA gene predictions and an interface for user predictions is available via the web.

INTRODUCTION

The value of genomic DNA sequence data is dependent on the algorithms and software available for interpretation and analyses of the nucleotide strings. For example, software packages such as Glimmer, GeneMark and GeneScan (1,2) are widely used to identify protein open reading frames (ORFs), while FASTA, BLAST and PSI-BLAST infer protein function from sequence homology. The ability to locate all or most protein ORFs in genomic sequences has led to the entire field of proteomics and the ability to monitor protein expression and function during cellular processes.

In contrast to the numerous successful programs and algorithms for identifying protein genes and coding sequences, virtually no computational methods are available for identifying novel genes for stable, functional RNAs (fRNAs) or regulatory elements in mRNA that control translation or stability. The major reason for this is that the signals used for finding protein genes, start and stop codons, the triplet amino acid code and ribosome-binding (Shine–Dalgarno) sequences, are not present in RNA genes. Other signals, such as promoters, terminators and processing sites, are not easily recognizable and thereby are unreliable indicators.

Currently, RNA genes are found in genomic sequences by their sequence or structural homology to known RNAs. Programs such as tRNAscan-SE, FAStRNA and Snoscan (3–5) utilize conserved elements of sequence and structure to identify tRNAs and snoRNAs. Genes for other known RNAs can often be located by sequence homology or motif searches (6). Folding free energy has been used to predict structured RNA elements as potential fRNAs (7,8). The Pol3scan (9) program uses a combination of polymerase III binding and terminator recognition sites together with base pairing motifs to find eukaryotic tRNAs. Identification of consensus polymerase III promoters has been used to guide an experimental search for new RNAs in yeast (10) that was successful in identifying several expressed RNAs. Recently, a comparative genomics study of the intergenic regions of *Escherichia coli* and other related genomes predicted the presence of 19 novel RNAs that were confirmed by biochemical experiment (11). Finally, an experimental study (12) has identified 201 potential new small RNAs expressed in mouse brain.

There is, however, no general computational approach to locating novel functional RNA sequences that lack sequence or structure homology to one of the limited number of known RNA types. Computational prediction of fRNAs in genomic sequences would allow experimental testing of expression levels, functional assay by deletion or mutagenesis, structural analysis and identification of protein or nucleic acid interaction partners. These untranslated fRNAs have also been referred to as non-coding RNA (ncRNA), small RNAs (sRNA, smRNA), untranslated and small non-messenger RNA (snmRNA). We will continue to refer to them as fRNA throughout this paper.

Our working hypothesis is that characteristic signals exist in the sequences of fRNAs that are distinguishable from, and in contrast to, sequences of non-coding regions of the genome. Non-coding regions are defined as those regions of the genome excluding protein genes and RNA genes, along with their promoters and terminators. We expect that the evolutionary forces responsible for the diversity of sequences of non-coding

*To whom correspondence should be addressed. Tel: +1 510 486 4305; Fax: +1 510 486 6059; Email: srholbrook@lbl.gov

Table 1. fRNAs in *E.coli*

RNA	Gene	Length	Function
tRNA	86 genes	~76 each	Translation of proteins
5S rRNA	8 gene copies	~120 each	rRNA, translation
16S rRNA	7 gene copies	~1540 each	rRNA, translation
23S rRNA	7 gene copies	~2900 each	rRNA, translation
MicF	*micF*	93	Antisense to *ompF*
DicF	*dicF*	53	Antisense to *ftsZ*
M1	*rnpB*	377	RNase P, RNA processing
4.5S	*ffs*	114	Protein translation, transport
DsrA	*dsr*	85	HNS antagonist, activator of *rpoS*
OxyS	*oxyS*	109	Activator/repressor, antimutator
10Sa	*ssrA*	363	tmRNA, protein tagging and degradation
CsrB	*csrB*	>360	CsrA antagonist
Spot 42	*spf*	109	Downstream of *polA*, inhibition of DNA synthesis
6S	*ssrS*	184	Regulator of RNA polymerase
RprA	*rprA*	101	Regulator of *rpoS*

Total nucleotides of RNA 41 697.

regions will lead to a different sequence distribution than that of either protein or RNA coding regions.

We propose to extract these signals from sequences of known fRNAs and non-coding regions using machine learning methods, specifically computational neural networks (NNs) and support vector machines (SVMs), and to apply these learned rules to the prediction of novel fRNAs among the currently unannotated regions of the genome. We present here our initial studies with prokaryotes and archaea, but the same approach is applicable to eukaryotes, including the human genome.

MATERIALS AND METHODS

Software

All neural network simulations and testing were done with the locally developed BIOPROP program (13). We have used the SVM^{light} program (14) for all support vector machine calculations. Large-scale calculations of the free energy of folding were made using a local program based on the Vienna RNA package (http://www.tbi.univie.ac.at/~ivo/RNA/) (15).

Genome sequences

Genome sequences and annotations were taken from their original sources. The *E.coli* sequence was from the *E.coli* genome project at the University of Wisconsin–Madison (http://www.genome.wisc.edu/). The genome sequences of *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Bacillus subtilis*, *Haemophilus influenzae* and *Deinococcus radiodurans* were from the TIGR microbial database (http://www.tigr.org/tdb), the *Pyrococcus horikoshii* genome sequence was from the National Institute of Technology and Evaluation (NITE), Tokyo, Japan (http://www.bio.nite.go.jp/ot3db_index.html) and the *Mycoplasma pneumoniae* sequence from the Heidelberg *M.pneumoniae* project (http://www.zmbh.uni-heidelberg.de/M_pneumoniae/genome/Results.html).

Compilation of sequence databases

We chose to formulate and test our predictive methods on the *E.coli* genome as the most well-studied bacterium in terms of both genome structure and function. The M52 version of the *E.coli* K-12 genome sequence (16) was used to compile a database of fRNA and non-annotated sequences. The well-characterized fRNA sequences of *E.coli* as compiled by Wassarman *et al*. (17) are shown in Table 1. The sequences of these RNA molecules served as positive examples from which we derived parameters for machine learning. The 'non-coding', or intergenic, sequences were obtained by removing all protein and known fRNA coding regions from the genome along with a buffer of 50 residues on both the 5′ and 3′ sides so as to remove possible promoter, terminator and other untranslated control elements. Sequences in both strands were removed when there was a protein or RNA coding region on either strand.

It is inherent in the nature of our problem, discovering possible RNA genes in non-annotated genomic DNA, that the 'non-coding' data will actually contain some real RNA genes. However, we assume that only a small fraction of the non-annotated sequences correspond to 'coding' sequences of RNA genes and we are therefore justified in using the non-annotated sequence as negative examples of RNA genes in machine learning. We make this assumption realizing that our non-coding database is somewhat contaminated with currently unknown RNA genes. After making our initial predictions, we filter our database in an iterative manner, removing strongly

predicted genes and retraining in order to 'purify' our non-coding database.

The complete *E.coli* genome consists of 4 697 221 nt on each strand (9 394 442 total). After removal of all protein coding genes, RNA coding genes (Table 1) and flanking regions, 337 662 nt remained on each strand for a total of 675 324 non-coding nucleotides. Our known RNA dataset consists of 41 697 nt. In order to remove bias from the database, duplicate RNA sequences were removed, leaving one copy of each rRNA and tRNA. This reduced the RNA database to 8400 nt. Thus, the non-coding database is ~80 times as large as the number of unique RNA coding nucleotides in *E.coli*.

Each RNA and non-annotated intergenic sequence was then divided into sequence windows of 80 residues with a 40 nt overlap between windows (i.e. each window slides 40 residues along the sequence). This window size was initially chosen to correspond to the size of tRNAs. Testing with different sized windows and overlaps showed that this choice was optimal for both prediction accuracy and computational speed. A window of <40 residues at the end of the sequence was omitted from the calculations. A total of 7705 windows from each strand (15 410 total) were partitioned from the non-coding sequences, while 188 unique RNA sequence windows were available from the known RNA sequences (after removing redundant RNAs). Of these 188 windows, 38.3% were from 23S rRNA, 26.6% from miscellaneous small RNAs, 20.2% from 16S rRNA, 13.3% from tRNA and 1.6% from 5S rRNA.

The large disparity in non-coding to RNA coding sequence windows presented a problem for neural network training, where the number of training examples of each type should be similar. One answer to this problem is to compile multiple datasets each having the same sequences from RNA genes, but with different non-coding sequences. To test this approach we made five datasets (*E.coli* 1–5) using the same 188 RNA coding windows but with a different unique set of 188 windows extracted randomly from the non-coding sequences. These five datasets were independently used to train and test neural networks and support vector machines and to predict novel RNA genes. An advantage of this approach is that predictions can be checked for agreement between the different networks, thus reducing false positives.

Selection and calculation of input parameters for machine learning

We have previously demonstrated that machine learning methods based on the composition, transition frequency and sequence distribution of amino acids can be used successfully in the prediction (18) of protein folding class from sequence alone. We therefore tested these parameters for their ability to discriminate sequence windows extracted from fRNAs from those arising from non-coding regions. Composition was represented as percent nucleotide composition, %A, %G, %C and %T (%U). Transition, represented in proteins as the alternation between hydrophobic and hydrophilic amino acids, was parameterized for RNA simply as the percentage of each dinucleotide present in the sequence window (i.e. %AA, %AG, etc.). Since we are using sequence windows rather than complete gene sequences as our prediction unit, the use of sequence distribution parameters does not apply.

As a supplement and complement to these compositional parameters we considered an additional set of parameters describing the occurrence frequency of sequence motifs commonly found as RNA structural elements. These included the well-known sequence motifs UNCG (19), GNRA (20) and CUYG (R, purine; Y, pyrimidine) found in RNA tetraloops (21,22) and the AAR (23) subsequence of the tetraloop receptor motif. In addition, the DNA sequence CTAG (RNA = CUAG) that occurs rarely in bacterial protein genes and non-coding regions compared to RNA genes (16,24) was included. Despite many studies and hypotheses regarding the basis of this sequence anomaly (25,26), the reasons for this bias are unclear, although it may have some structural basis (27). The final parameter of this set was the calculated free energy of folding for the RNA sequence window. This parameter was chosen based on calculations showing the average calculated free energy of folding (28) of the sequence windows corresponding to known RNAs (*E.coli*, -2.70 ± 0.52 kcal/80 nt; *M.jannaschii*, -3.68 ± 0.72) to be lower than that calculated from the non-coding sequence windows (*E.coli*, -2.06 ± 0.85; *M.jannaschii*, -1.34 ± 0.66).

The twenty 'compositional' parameters (four for percent nucleotide composition, 16 for percent dinucleotide composition or transition) and the six 'structural motif' parameters described above were calculated for all sequence windows and used in training and testing of neural networks and support vector machines and for prediction of novel RNA genes by the trained computational machines.

Neural network architecture

All neural networks used in training, testing and prediction were of the back-propagation, feed-forward type with a single hidden layer. The 20 input 'composition' parameters were used for testing and training of a neural network with three hidden nodes, while a separate neural network of similar architecture was trained and tested using the six 'structural motif' parameters.

In order to optimize prediction, the predictions (output activities) from the composition and structural motif networks were used as input to a third neural network (voting network), which is trained to make a final decision as to whether a sequence window belongs to a RNA gene. The overall network architecture is shown in Figure 1.

RESULTS AND DISCUSSION

Neural networks for RNA gene prediction in *E.coli*

Networks were tested by a full jackknife procedure (removing one example at a time, retraining, testing on the single example, then averaging over all examples). The results of the jackknife testing experiments for *E.coli* (dataset 1) are shown in Table 2. These results are evaluated in terms of the contingency matrix (true and false positives and negatives), the correlation coefficient and Q^{α} (average of the percentage of correctly predicted positive windows and the percentage of correctly predicted negative windows) (29).

The composition network has a prediction accuracy, Q^{α}, of >85%, while the structural motif network alone predicts RNA genes with slightly over 81% accuracy. However, the voting network, which takes into account both composition and structural parameters, performs best, with 92% accuracy over all examples, rising to 93.5% for the strongest predictions.

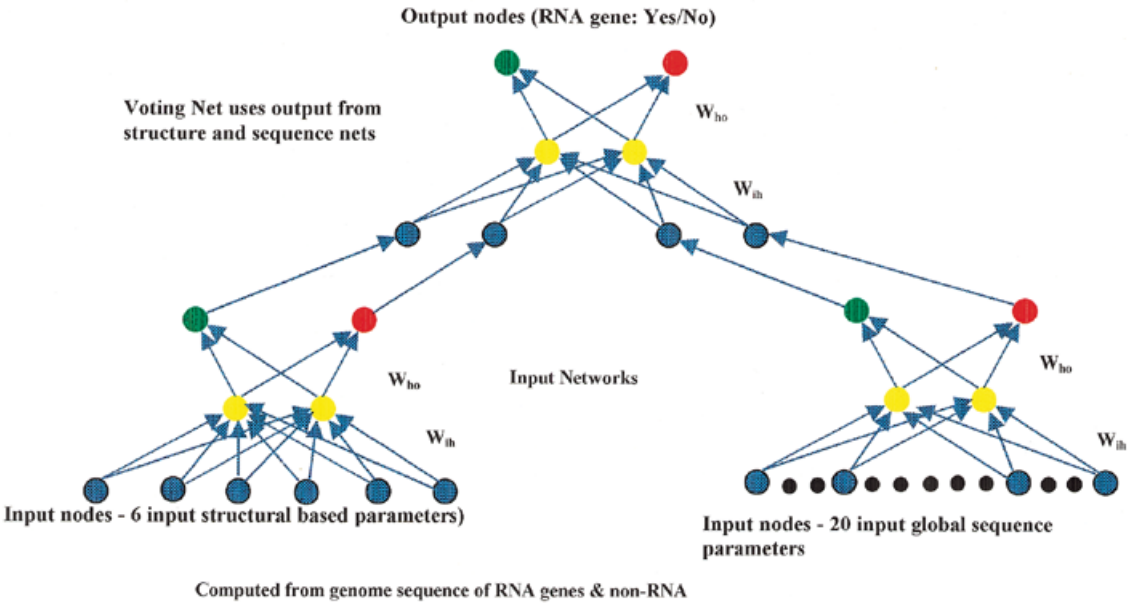

Figure 1. Architecture of computational neural networks used for the prediction of fRNAs. Input parameters are supplied to structural motif (six input nodes) and compositional (20 input nodes) neural networks as described in the text. Each of these networks utilizes a middle layer of hidden nodes (three hidden nodes in each network). The output activities from the structural and compositional networks are used as input to the voting network (one hidden node) that makes the ultimate prediction as to whether the sequence window is part of a fRNA. The strength of the links between nodes are the weights (W_{ij}) indicated in the figure.

Table 2. Testing of neural networks for prediction of RNA genes in *E.coli*

Threshold	Structural motif network						Compositional network						Voting network					
	TP	FP	FN	TN	Q^{α}	CC	TP	FP	FN	TN	Q^{α}	CC	TP	FP	FN	TN	Q^{α}	CC
0.5	163	45	26	142	81.1	0.625	157	23	32	165	85.4	0.709	171	13	17	176	92.0	0.841
0.6	154	38	20	132	83.1	0.666	154	22	27	163	86.6	0.732	171	11	17	173	92.5	0.850
0.7	136	31	15	126	85.2	0.706	154	21	23	160	87.7	0.754	171	11	17	173	92.5	0.850
0.8	87	17	14	114	86.6	0.729	149	17	19	149	89.2	0.784	171	11	17	171	92.5	0.849
0.9	24	4	7	101	86.8	0.764	132	15	13	140	90.7	0.813	168	11	11	148	93.5	0.869

TP, true positive; FP, false positive; FN, false negative; TN, true negative; Q^{α}, average percent correct as defined in Baldi *et al.* (29); CC, correlation coefficient (29).

Notably, at least one window of every RNA in our dataset was predicted in testing experiments.

We also tested a further four *E.coli* datasets (each with the same RNA examples and a different set of non-coding examples) using the same jackknife procedure. The average Q^{α} for the composition network was 82.9%, whilst that for the structural motif network was 75.7%. The voting network has an average Q^{α} of 89.4%. The same pattern is seen in the correlation coefficients (Table 3), with the voting network having the highest correlation (average over five datasets, 0.789), followed by the composition network (correlation coefficient 0.658), with the weakest correlation for the network based on structural motifs (correlation 0.522).

To evaluate the effect of dataset contamination (RNA sequences within the non-coding data) in the *E.coli* datasets, any negative windows that were strongly predicted to be positive in both the structural and compositional nets were removed and the networks retrained. Removal of 10 examples from *E.coli* dataset 1 and 14 examples from dataset 3 led to improvements in the final testing accuracy of ~2.2 and 1.6%, respectively. These examples were then tested by both retrained networks from which they had been omitted and all 24 examples were predicted to be RNA genes by both networks.

To further examine the effect of contamination, two additional experiments were undertaken. In the first, any window that was predicted positive by any of the trained networks was removed from the negative set. Retraining using the same positives and a random set of these 'absolute negatives' gave a significant improvement in Q^{α} (96.8% compared to the

Table 3. Compiled neural network results for all organisms

Data	Training net	No. of input windows	TP	FP	FN	TN	Q^{α}	Correlation coefficient
E.coli 1	4	376	171	13	17	176	92.0	0.841
	6		163	45	26	142	81.1	0.625
	20		157	23	32	165	85.4	0.709
E.coli 2	4	376	161	15	27	173	88.8	0.778
	6		157	75	31	112	71.7	0.447
	20		144	33	44	153	79.4	0.589
E.coli 3	4	376	169	18	19	170	90.2	0.803
	6		160	58	28	130	77.1	0.550
	20	376	158	28	31	160	84.4	0.687
E.coli 4	4	376	160	18	28	171	87.8	0.757
	6		151	56	37	132	75.3	0.508
	20		153	35	35	153	81.4	0.628
E.coli 5	4	376	176	33	12	155	88.0	0.765
	6		160	72	28	116	73.4	0.482
	20		153	26	35	161	83.7	0.675
B.subtilis	4	324	153	0	9	162	97.2	0.946
	6		140	30	22	132	84.0	0.680
	20		150	14	12	148	92.5	0.840
M.genitalium	4	312	147	23	9	133	89.7	0.798
	6		126	41	31	114	76.9	0.539
	20		131	31	24	125	82.3	0.647
M.pneumoniae	4	318	145	11	14	148	92.1	0.843
	6		128	48	31	111	75.2	0.506
	20		133	25	26	134	84.0	0.679
H.influenzae	4	270	133	3	2	132	98.1	0.963
	6		118	27	17	109	83.8	0.677
	20		124	9	11	127	92.6	0.852
D.radiodurans	4	224	105	18	7	94	88.8	0.781
	6		83	46	29	66	66.5	0.334
	20		91	21	21	91	81.2	0.625
M.jannaschii	4	278	139	1	0	138	99.6	0.993
	6		133	8	6	131	95.0	0.899
	20		138	2	1	137	98.9	0.978
P.horikoshii	4	322	148	5	13	156	94.4	0.889
	6		138	16	23	146	87.9	0.759
	20		140	15	20	146	89.1	0.782
Bacteria	4	1500	657	86	94	664	88.0	0.760
Bacteria except *D.radiodurans*	4	1276	574	77	63	561	89.0	0.781
Archeae	4	600	285	11	15	289	95.7	0.913

previous 92.0%) and, as expected, there were no false positives.

In the second experiment all of the negative windows used in training sets 1–5 were searched for homology using BLAST (30). Any window that showed any homology to any other sequence in the NR database was removed from the negative set (homology was determined as any window that had a BLAST expectation value $<10^{-3}$). When this network was

Table 4. Known fRNAs in *E.coli* predicted with RNAGENiE

Gene (size)	Reference	Prediction	Function
crp (94 nt)	(34)	Predicted (1/2 windows)	Expression regulation
RQ120 (120 nt)	(35)	Predicted (1/2 windows)	Recombinant RNA
029A small RNA (101 nt)	(44)	Predicted (1/2 windows)	Unknown
029B small RNA (118 nt)	(44)	Not predicted	Unknown
CopA (91 nt)	(45)	Not predicted	Expression inhibition
Plasmid 1162 (115 nt)	(36)	Predicted (2/2 windows)	Expression inhibition
gcvB RNA (265 nt)	(46)	Not predicted	Gene repression
tyrT	(32)	Predicted	Modulator
QUAD	(33)	Predicted (5/5 windows)	Unknown
PAIR	(33)	Predicted (2/3 windows)	Unknown

retrained again an improvement in Q^{α} was seen (95.5%) and a reduction in the number of false positives was noticed (8 compared to 15).

The high prediction accuracy for each of the networks and datasets illustrates that neural networks are able to learn to distinguish between RNA genes and non-coding regions with high accuracy using either structural or compositional parameters as input. It is also apparent that a voting network incorporating output from both structural and global composition networks is the most accurate and predicts the greatest percent of the data at all threshold levels. We thus feel justified in using a voting network trained with the architecture and parameters described above to search the intergenic regions of *E.coli* for novel fRNAs.

Neural network prediction results for *E.coli*

Using the trained voting network from each *E.coli* network and compiling the results we found that 285 sequence windows were predicted in all networks to belong to RNA genes in strand 1 and 277 windows to RNA genes in strand 2, out of the 15 410 intergenic sequence windows in each strand of *E.coli* (1.8%). Since many of these predicted windows are consecutive, the actual number of RNA genes predicted is smaller. If consecutive windows and cross-strand predictions are taken into account, approximately 370 novel fRNAs are predicted in the *E.coli* genome.

It was also noted that a number of the predicted RNA genes have strong sequence homology to regions in the *E.coli* O157:H7 genome (31) as determined using a BLAST search (homology was determined as any window that had a BLAST expectation value $<10^{-5}$). The non-coding sequence between bases 2151201 and 2151760 is analogous to a region of the O157:H7 genome over 523 of 546 bases and is strongly predicted over four windows in this region. Many predicted genes were also homologous to regions of *Salmonella typhimurium* and other bacterial genomes closely related to *E.coli.* The putative gene between bases 4177162 and 4177523, which is strongly predicted over three windows, is identical for a 50 nt section to a region of the O157:H7 genome; it is also conserved for 55 of 56 nt in *Citrobacter freundii* and 52 of 54 nt of *S.typhimurium*. It is part of a ribosomal protein operon in both *E.coli* and *C.freundii*.

Of the putative *E.coli* RNA genes predicted by the trained networks, many clustered either together or with known RNAs. A striking example is observed in the spacer regions between tRNA genes found in several tRNA operons. In one operon consisting of lysine and valine tRNA genes, each of five intergenic regions is predicted as an RNA gene. These spacers are similar in sequence, as shown by multiple sequence alignment, can fold into a similar secondary structure and have homologs in other *E.coli* tRNA operons and in related organisms such as *S.typhimurium*.

A search of the literature revealed a number of fRNAs that were not included in the original training set. These are summarized in Table 4. A small RNA of 171 nt from the *tyrT* operon of *E.coli* has been experimentally identified (32) and proposed to have a modulatory effect on stringent response. The sequence windows corresponding to this RNA were strongly predicted by all *E.coli* networks to be a fRNA. A number of the known DNA repeat sequences in the *E.coli* genome are also predicted by the neural networks to be RNA genes. The predictions included the TRIP repeat, a 266 nt sequence, and the QUAD repeat, a 165 nt sequence, both previously speculated to be RNA genes (33).

The *crp* divergent RNA gene (34), which regulates expression, was identified as a 94 nt RNA and was successfully predicted. RQ120, a 120 nt RNA that was identified as a recombinant RNA produced in *E.coli* in the presence of Qβ replicase (35) was successfully identified. Finally, a small RNA that inhibits expression of the *RepI* gene (36) of plasmid R1162 was also successfully identified by our networks. Further examples of successfully predicted known fRNAs are indicated in Table 4. In future iterations of RNAGENiE these examples can be used in the training set to improve accuracy.

Using comparative genomics to identify conserved intergenic sequences between *E.coli* and related genomes, Wassarman *et al.* (11) have postulated the presence of 26 potential new RNA genes. Twenty-five expressed RNAs corresponding to these sequences were verified by northern blot and microarray analysis to be present in *E.coli*. Nineteen of these RNAs appear to be non-messenger while the remaining six appeared to code for putative ORFs. Using RNAGENiE we correctly predicted 17 of these 19 novel RNA genes (~90%) with at least one sequence window. One site of disagreement corresponds to a

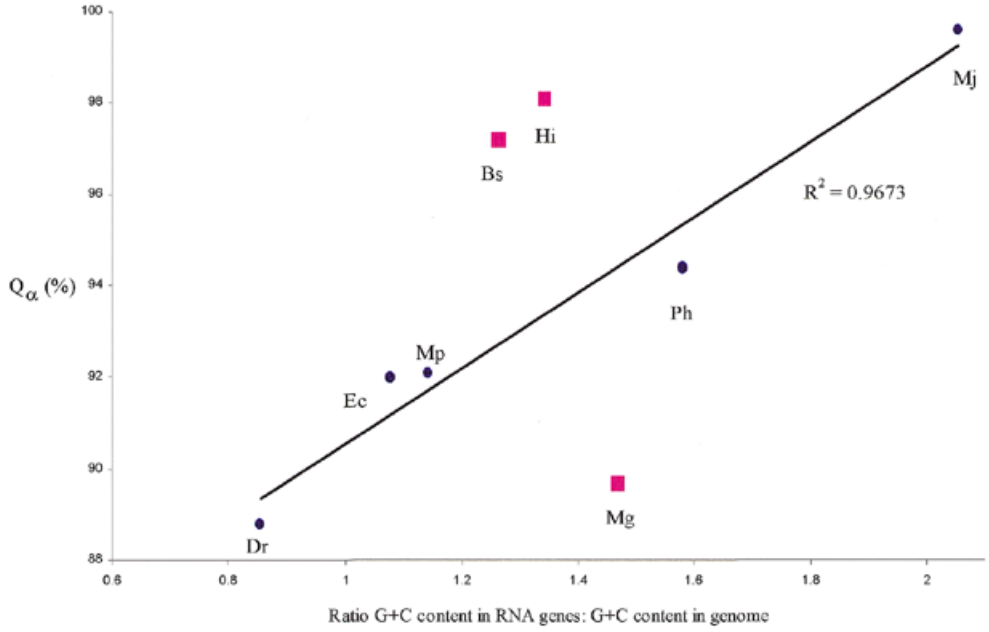

Figure 2. Q^α plotted against the ratio of G+C content in the fRNAs to the overall genomic G+C content. The trend line and correlation coefficient are shown for *D.radiodurans* (Dr), *E.coli* (Ec), *M.pneumoniae* (Mp), *P.horikoshii* (Ph) and *M.jannaschii* (Mj). There is a strong correlation between the G+C content for these organisms and the ability of RNAGENiE to predict fRNA. For the three remaining organisms that were examined, *B.subtilis* (Bs), *H.influenzae* (Hi) and *M.genitalium* (Mg), the correlation is not so strong. In the majority of cases RNAGENiE performs better than expected due to the influence of additional parameters to composition.

region in which RNAs were predicted and found to be partially overlapping on opposite strands of the genome. While we correctly predict an RNA coding region on the Watson strand we do not find any signal on the Crick strand. Finally, another intergenic region was found by Wassarman *et al.* (11) to contain two adjacent RNA genes, while we predict only the first one. It is possible that in this case there is only a single larger expressed RNA that appears as two sequences after processing.

Training and testing neural networks in other genomes

We have applied the prediction scheme described above to and tested it on several other organisms, including the bacteria *B.subtilis* (Bs), *M.genitalium* (Mg), *M.pneumoniae* (Mp) *D.radiodurans* (Dr) and *H.influenzae* (Hi) and the hyperthermophilic archaea *M.jannaschii* (Mj) and *P.horikoshii* (Ph) (Table 3).

The best performance on jackknife testing of the prokaryotes was for Bs and Hi, which had Q^α scores of 97.2 and 98.1%, respectively. These two genomes have a relatively high G+C content in their RNA gene sequences compared to the overall G+C content in their genomes, 1.26 and 1.34, respectively. The Dr genome, which performed least well on testing (88.8% correct), had the lowest G+C ratio of all genomes tested, 0.85.

Mg and Mp have small and highly related genomes, with the Mg genome almost a subset of Mp. Testing of neural networks for RNA prediction in Mg gave a Q^α of 90% and for Mp 92%. When the network trained on Mg was used to predict Mp RNAs, it was 89.3% accurate, while the Mp networks were 89.7% accurate in predicting Mg RNAs. This cross-prediction between organisms suggests that we may be able to predict RNA genes in many organisms from related species rather than having to make individual networks for each organism.

The prediction of RNAs in Mj and Ph, both hyperthermophilic archaeal organisms, was extremely accurate, with Q^α values of 99.6 and 94.4%, respectively. As observed for prokaryotes, this accuracy may partly arise from the relatively high overall G+C content of RNAs (37) in these hyperthermophiles relative to the rest of the genome. The ratio of G+C content in the RNA genes to overall genomic G+C content for Mj is 2.05 and for Ph 1.58. To further examine this relationship, we show in Figure 2 a plot of Q^α versus G+C ratio for all genomes we have studied. As can be seen, there is a high degree of correlation for most organisms, indicating, as suggested by Rivas and Eddy (37), that high G+C content is a key indicator of RNA genes. However, three genomes are well off the line, with Bs and Hi having a higher prediction correlation than expected and Mg lower than expected based strictly on G+C content. Clearly, other factors than composition must be taken into account to successfully discriminate between RNA genes and non-coding DNA in these organisms.

RNA gene prediction in other prokaryotes and archaea

After verification of neural network performance, predictions are made for all intergenic sequence windows of each genome as to whether they code for a fRNA. Each prediction is associated with an activity or strength of prediction that is related to the likelihood that the prediction is correct. Thus some sequences are more confidently predicted than other sequences. Consecutive predicted sequences define an RNA that is larger than the sequence window of 80 residues. We have applied this procedure for the prediction of fRNAs for each of the genomes listed

Table 5. Other known fRNAs predicted with RNAGENiE

Organism	Gene (size)	Reference	Prediction	Function
M.pnuemoniae	*Mp200* (212 nt)	(38)	Predicted (2/5 windows)	Reducing agent?
M.genitalium	*Mg170* (170 nt)	(38)	Predicted (2/4 windows)	Reducing agent?
D.radiodurans	*Ro* gene a (146 nt)	(39)	Predicted (1/3 windows)	UV repair
	Ro gene b (145 nt)	(39)	Not predicted	UV repair
	Ro gene c (126 nt)	(39)	Predicted (1/3 windows)	YRNA
	Ro gene d (96 nt)	(39)	Predicted (1/3 windows)	UV repair
M.jannnaschii	SnoRNA (~56 nt each)	(40)	Not predicted (1 of 8 predicted)	rRNA modification

and identified many putative fRNAs. These predicted RNAs are stored in a database on our web site. As in *E.coli*, several predicted RNAs not included in our original database have been identified and characterized in the literature. These are summarized in Table 5 and described below.

A new small RNA gene (200 bp) was recently discovered in Mp, along with a homolog in Mg (38). These RNAs were not included in our training set. The RNAs are expressed with high copy number and due to its conservation between Mp and Mg, the authors speculate that it is likely to be a fRNA. The trained neural networks for each organism successfully predicted the corresponding gene, with each network strongly predicting the most highly conserved region of the putative RNA sequences.

Four small RNAs, also not included in our database, were found to be expressed from the Dr genome in response to UV radiation (39). These RNAs bind to the product of the *rsr* gene (Ro autoantigen ortholog) from *D.radiodurans*. One of these RNAs was identified as having structural and sequence similarities to eukaryotic YRNAs. When these sequences were tested, it was found that three of the four RNAs had at least one sequence window predicted as a fRNA.

snoRNAs have recently been identified in Mj and other archaea (40) using the Snoscan program. Our program fared reasonably well with these predicted guide RNAs that anneal to rRNA. Each snoRNA is only ~56 nt in length and so is just a single sequence window. Out of a total of 59 snoRNAs in Mj and Ph our program predicted 35. Their small size and the fact that snoRNAs require rRNA as a template to fold into their functional form suggests that special paramaterization may be necessary to predict these types of RNAs.

Support vector machines

An alternative approach to classification problems by machine learning uses the so-called support vector machine (SVM) (41). Using the structural motif-based and compositional parameters described above as input, we created a combined input pattern file from which we have calculated SVMs to classify the sequence windows as to whether they are part of fRNAs or not. The results of this classification are shown in Table 6 for all organisms. These results were obtained using either a third degree polynomial or a radial basal kernel function and are all comparable to, although somewhat less accurate than, the neural network results.

Table 6. SVM results

Organism	Kernel[a]	L-O-O estimate of error[b]	Recall[c]	Precision[d]
E.coli 1	RBE	12.0	89.9	86.7
E.coli 2	RBE	16.8	86.7	81.1
E.coli 3	3°	13.1	90.9	84.1
E.coli 4	RBE	14.6	87.2	84.1
E.coli 5	RBE	14.4	88.8	83.5
B.subtilis	3°	6.5	95.7	91.7
M.genitalium	RBE	12.8	87.2	87.2
M.pneumoniae	3°	12.6	88.7	86.0
H.influenzae	RBE	5.9	98.5	90.5
D.radiodurans	RBE	16.3	87.5	81.0
M.jannaschii	3°	0.7	100.0	98.6
P.horikoshii	RBE	8.1	87.6	95.9

[a]Kernel function used in SVM: RBE, radial basal equation; 3°, third degree polynomial.
[b]Similar to neural network jackknife testing, in which selected examples are removed, the SVM retrained and a prediction made for the removed example. Estimated error is the average of all of these tests.
[c]Recall is the probability that a known example (training example) is classified correctly.
[d]Precision is the probability that an unknown example (testing example) is classified correctly.

Contributions of neural network input parameters to prediction

The weights linking nodes in the computational neural networks can be analyzed to assess which input parameters are most important in making a prediction of a sequence window as belonging to an RNA gene or not. In this manner, a physical basis for the predictions may be discovered and less informative parameters can be removed. In order to simplify this analysis, the networks were retrained as perceptrons, with no hidden nodes. We have conducted such an analysis for two *E.coli* training sets using the known RNA dataset and two different negative datasets (*E.coli* 1 and *E.coli* 2).

For the composition network, the nucleotide compositions of G and T are most significant, as indicated by the weights, with

%G favoring and %T disfavoring prediction as an RNA gene. Likewise, the dinucleotide compositions most important are %CT, %GT and %GG, which support RNA gene prediction, and %GC and %AT, which oppose it.

Analysis of the structural motif network showed that the largest weights were for links corresponding to the GNRA and CUAG occurrence frequency and the calculated free energy of folding. Increases in each of these parameters (more negative free energy of folding) favors prediction as an RNA gene. The UNCG, AAR (tetraloop receptor) and CUYG sequence frequencies are associated with low weights and are therefore less important for prediction.

Finally, the voting net, in agreement with the accuracy of the individual structure net and composition net, much more strongly weights the composition network (greater than 2:1) relative to the structure network.

To further analyze the effect of individual parameters, the networks were retrained leaving out each parameter in turn from both the structural motif and compositional networks. The results from these modified networks were fed forward into a voting net that was also retrained. The most significant observation was that removal of the free energy term had a major effect on predictive accuracy of the structural motif network alone, lowering Q^{α} by >5%. However, this omission did not affect the results of the voting network significantly, reflecting the robustness of the parameters.

To ensure that the neural networks were not being trained to recognize just tRNA or rRNA (the majority of the training set) but were instead recognizing some broad general parameters of all RNA genes, cross-prediction experiments were undertaken. In three experiments, either all the tRNA windows, all the windows from 23S rRNA or all the windows from the small fRNA genes were excluded from training and then the trained weights were used to predict the excluded windows. In all three cases the results supported generalization of the trained networks. In the case of tRNA, 21 tRNAs were excluded (corresponding to 25 windows). At the lowest threshold 18 of the 21 tRNAs were predicted (21 of 25 windows, 84% correct). For the 23S rRNA, 72 windows were excluded and from the trained set 62 windows were predicted to code for RNA (86% correct).

Analysis of the predictions with the other small RNAs excluded showed that six of 11 excluded RNAs (26 of the 50 windows) were predicted. This shows that there is cross-prediction between different classes of fRNAs. However, including all the known small RNAs in the input data trains the neural nets on as diverse a dataset as possible, allowing it to recognize new members of the fRNA family. This is shown by an increase in the prediction level from 50% when all small RNAs are excluded up to the 88% seen in the full input set.

Rivas and Eddy (37) state that the calculated stability of most fRNA secondary structures is not sufficiently different from the predicted stability of a random sequence of the same composition to be useful as an RNA gene finding approach. However, if tRNAs are excluded, these authors also show that fRNAs generally do have higher thermodynamic stabilities compared to random sequences, but that this difference is small. Our approach does not compare calculated stability of fRNAs to random sequences, but rather to non-coding sequences that have biases of their own. Our results indicate that together with sequence motifs known to occur commonly in RNA molecules, the calculated free energy of folding can improve prediction of known RNAs when compared to the use of composition alone. We agree with Rivas and Eddy that base composition is the key factor in distinguishing RNA genes from non-coding sequences, but our results suggest that in some cases RNAs can be identified using structural parameters, including free energy of folding.

Combined species prediction

To test the species-specific bias of our predictive approach, a number of networks were trained using combined data from different organisms. The results are shown in Table 3. One set consisting of both sets of archaeal data (Mj + Ph) performed very well upon jackknife testing, having a Q^{α} of 96%. The two trials using combined sets of bacterial data performed less well, with $Q^{\alpha} < 90\%$. This indicates that common features may describe and distinguish RNA genes for the hyperthermophilic archaea, but distantly related bacteria may encode RNA genes with different characteristics that require species-specific networks for discrimination. Further experiments with more genomes are necessary to definitively answer this question.

Prediction of fRNA regions in mRNA

The success discussed above in predicting RNA genes suggested the possibility that the same or similar algorithms could be used to predict other fRNA regions, such as the untranslated control regions (UTRs) of some mRNAs that are also generally highly structured. We tested this idea on an evolutionarily conserved RNA stem–loop in the 5′-UTR of RNase E mRNA (42). This regulatory element was strongly predicted by our method, with the boundaries well defined by the prediction.

We have also tested the *cis*-acting mRNA element called the selenocysteine insertion sequence (SECIS) occurring in the *E.coli fdhF* gene mRNA (43). This 42 nt sequence, which forms a stem–loop structure, was strongly predicted as a fRNA region by our program.

Clearly, extensive testing remains to be done to determine whether the same or related prediction methods can be used to identify all fRNAs, be they individual genes or control regions of mRNA.

CONCLUSIONS

The results described above clearly show that RNA genes can be identified with high confidence in bacterial and archaeal genomes using a machine learning approach based on differences in compositional and structural parameters present in known RNAs compared to non-coding sequences. In addition to cross-validation testing, strongly predicted sequences in *E.coli*, *M.genitalium*, *M.pnuemoniae* and *D.radiodurans*, not included in the training datasets, have been experimentally characterized and reported in the literature as expressed fRNAs. Further credibility is given to our predictions by the observation that these sequences are highly conserved among closely related genomes, whereas non-coding sequences are not. We also make the preliminary observation that control regions in untranslated mRNA can be identified as fRNAs. Further studies are necessary to characterize the extent of these elements and the accuracy of their prediction.

Although many parameters are important in assignment of a sequence as part of an RNA gene, one major discriminator of genes appears to be the local G+C ratio. Thus, prediction accuracy is greater in genomes, such as the hyperthermophiles Mj and Ph, which have a much higher G+C composition in the RNA genes than in the overall genome. This bias is likely due to their requirement for heat-stable structures in their fRNAs.

Analysis of the input patterns from the fRNA using cluster analysis showed that the patterns did not contain any discernible clusters. This indicates that we have not produced a machine learning method that mimics certain classes or families of fRNAs but rather that we have trained our learning machines to recognize certain characteristics that are inherent in all fRNAs. However, our experience with the archeal snoRNAs shows that there are still some classes which contain signals sufficiently different from the general class that they are not consistently recognized. However, as more fRNAs are identified the learning machines will be trained on a wider variety of fRNAs and will thus be able to recognize broader parameters in fRNAs and so better encompass outliers.

We have begun studies with a variety of other parameters to optimize RNA gene prediction in prokaryotes and archaea. We are also extending this approach to the identification of novel fRNAs in eukaryotic organisms such as yeast, *Caenorhabditis elegans* and human. Although these organisms will have a larger database of known RNAs for machine learning, additional complications such as the intron/exon structure of genes must be considered.

Availability

These programs are implemented and accessible on the World Wide Web through the RNAGENiE interface (http://rnagene.lbl.gov/). Here users type or paste a sequence of interest into a text window. The program then divides the data into windows of the appropriate size, converts them to the proper input parameters and performs a prediction using the stored neural network weights. The results from RNAGENiE are then emailed to the users giving scores corresponding to the likelihood of any window being part of an RNA gene.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Chris Mayor in program and web site development and Chris Ding for his assistance and computational tests using support vector machines. Thanks are also due to Richard Meraz for critical reading of this manuscript and advice on machine learning methods and Adam Arkin for use of his program for cluster analysis. The authors also acknowledge Lawrence Berkeley National Laboratory and the Department of Energy for early support of this research through the Laboratory Directed Research and Development Program.

REFERENCES

1. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4341.
2. Ramakrishna,R. and Srinivasan,R. (1999) Gene identification in bacterial and organellar genomes using GeneScan. *Comput. Chem.*, **23**, 165–174.
3. El-Mabrouk,N. and Lisacek,F. (1996) Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome. *J. Mol. Biol.*, **264**, 46–55.
4. Lowe,T. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
5. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
6. Bourdeau,V., Ferbeyre,G., Pageau,M., Paquin,B. and Cedegren,R. (1999) The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.*, **27**, 4457–4467.
7. Le,S.V., Chen,H., Currey,K.M. and Maizel,J.V.,.Jr (1988) A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, **4**, 153–159.
8. Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, **27**, 1578–1584.
9. Percudani,R., Pavesi,A. and Ottonello,S. (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **268**, 322–330.
10. Olivas,W.M., Muhlrad,D. and Parker,R. (1997) Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.*, **25**, 4619–4625.
11. Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.
12. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.
13. Holbrook,S.R., Muskal,S.M. and Kim,S.-H. (1992) Predicting protein structural features with artificial neural networks. In Hunter,L. (ed.), *Artificial Intelligence and Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 161–194.
14. Joachims,T. (1998) Making large-scale support vector machine learning practical. In Scholkopf,B., Burges,C.J.C. and Mika,S. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.
15. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
16. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K. and Mayhew,G.F. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
17. Wassarman,K.M., Zhang,A. and Storz,G. (1999) Small RNAs in *Escherichia coli*. *Trends Microbiol.*, **7**, 37–45.
18. Dubchak,I., Muchnik,I., Holbrook,S.R. and Kim,S.-H. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA*, **92**, 8700–8704.
19. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
20. Jucker,F.M., Heus,H.A., Yip,P.F., Moors,E.H.M. and Pardi,A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, **264**, 968–980.
21. Woese,C.R., Winker,S. and Gutell,R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl Acad. Sci. USA*, **87**, 8467–8471.
22. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
23. Costa,M. and Michel,F. (1997) Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution. *EMBO J.*, **16**, 3289–3302.
24. Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.
25. Bhagwat,A.S. and McClelland,M. (1992) DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Res.*, **20**, 1663–1668.
26. Gutierrez,G., Casadesus,J., Oliver,J.L. and Marin,A. (1994) Compositional heterogeneity of the *E. coli* genome: a role for VSP repair? *J. Mol. Evol.*, **39**, 340–346.

27. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
28. Zuker,M. (1989) Computer prediction of RNA structure. *Methods Enzymol.*, **180**, 262–288.
29. Baldi,P., Brunak,S., Chauvin,Y. andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
31. Perna,N.T., Plunkett,G., Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al*. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
32. Bosl,M. and Kersten,H. (1991) A novel RNA product of the tyrT operon of *E.coli*. *Nucleic Acids Res.*, **19**, 5863–5870.
33. Rudd,K.E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.*, **150**, 653–664.
34. Okamoto,K., Hara,S., Bhasin,R. and Freundlich,M. (1988) Evidence *in vivo* for autogenous control of the cyclic AMP receptor protein gene (crp) in *Escherichia coli* by divergent RNA. *J. Bacteriol.*, **170**, 5076–5079.
35. Munishkin,A.V., Voronin,L.A. and Chetverin,A.B. (1988) An *in vivo* recombinant RNA capable of autocatalytic synthesis by Q beta replicase. *Nature*, **333**, 473–475.
36. Kim,K. and Meyer,R.J. (1986) Copy-number of broad host-range plasmid R1162 is regulated by a small RNA. *Nucleic Acids Res.*, **14**, 8027–8046.
37. Rivas,E. and Eddy,S. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
38. Gohlmann,H.W.H., Weiner,J., Schon,A. and Herrmann,R. (2000) Identification of a small RNA within the pdh gene cluster of *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *J. Bacteriol*., **182**, 3281–3284.
39. Chen,X., Quinn,A.M. and Wolin,S.L. (2000) Ro ribonucleoproteins contribute to the resistance of *Deinococcus radiodurans* to ultraviolet irradiation. *Genes Dev*., **14**, 777–782.
40. Omer,A.D., Lowe,T.M., Russell,A.G., Ebhardt,H., Eddy,S.R. and Dennis,P.P. (2000) Homologues of small nucleolar RNAs in Archaea. *Science*, **288**, 517–522.
41. Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK.
42. Diwa,A., Bricker,A.L., Jain,C. and Belasco,J.G. (2000) An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression. *Genes Dev.*, **14**, 1249–1260.
43. Li,C., Reches,M. and Engelberg-Kulka,H. (2000) The bulged nucleotide in the *E.coli* minimal selenocysteine insertion sequence participates in interaction with SelB: a genetic approach. *J. Bacteriol*., **182**, 6302–6307.
44. Guo,P.X., Bailey,S., Bodley,J.W. and Anderson,D. (1987) Characterization of the small RNA of the bacteriophage phi29 DNA packaging machine. *Nucleic Acids Res.*, **15**, 7081–7090.
45. Gerhart,E., Wagner,H. and Nordstrom,K. (1986) Structural analysis of an RNA molecule involved in replication control of plasmid R1. *Nucleic Acids Res.*, **14**, 2523–2538.
46. Urbanowski,M.L., Stauffer,L.T. and Stauffer,G.V. (2000) The gcvB gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol. Microbiol.*, **37**, 856–868.