Nucleic Acids Research, 2005, Vol. 33, No. 10 ***3263–3270***
doi:10.1093/nar/gki644

Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming

Pål Sætrom*, Ragnhild Sneve[1], Knut I. Kristiansen[1], Ola Snøve Jr., Thomas Grünfeld, Torbjørn Rognes[1] and Erling Seeberg[1]

Interagon AS, Medisinsk teknisk senter, NO-7489 Trondheim, Norway and [1]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Rikshospitalet University Hospital, NO-0027 Oslo, Norway

Received November 24, 2004; Revised February 22, 2005; Accepted May 20, 2005

ABSTRACT

Several methods exist for predicting non-coding RNA (ncRNA) genes in *Escherichia coli* (*E.coli*). In addition to about sixty known ncRNA genes excluding tRNAs and rRNAs, various methods have predicted more than thousand ncRNA genes, but only 95 of these candidates were confirmed by more than one study. Here, we introduce a new method that uses automatic discovery of sequence patterns to predict ncRNA genes. The method predicts 135 novel candidates. In addition, the method predicts 152 genes that overlap with predictions in the literature. We test sixteen predictions experimentally, and show that twelve of these are actual ncRNA transcripts. Six of the twelve verified candidates were novel predictions. The relatively high confirmation rate indicates that many of the untested novel predictions are also ncRNAs, and we therefore speculate that *E.coli* contains more ncRNA genes than previously estimated.

INTRODUCTION

Non-coding RNAs (ncRNA) are transcripts, whose function lies in the RNA sequence itself and not as information carriers for protein synthesis. Although long believed to be a minor gene class, recent discoveries have revealed that ncRNA genes are far more prevalent than previously believed and that they have other important roles beyond protein synthesis (rRNA and tRNA) (1–5).

In *Escherichia coli*, the number of experimentally verified small RNA (sRNA) genes (ncRNA genes excluding rRNA and tRNA) has increased rapidly. Only 10 sRNA genes were known in 1999 (6), whereas a recent survey listed 55 known sRNA genes (7). Subsequent RNA cloning experiments increased the number of known sRNA genes to 62 (8).

Most of these sRNA genes were identified in six studies describing systematic searches for new sRNA genes (9–14). All but one of these studies (14) used computational methods to predict sRNA genes. The computational methods ranged from analysis of sequence (9,10) and structure (11) conservation; to promoter and terminator identification (9,13); and machine learning based on sequence composition, known ncRNA motifs and RNA secondary structure stability (12). Together, these six studies have predicted ~1000 non-redundant sRNA candidates that are yet to be confirmed (7). Note, however, that only 95 candidates were predicted by more than one study.

We describe a method that uses automatic discovery of sequence patterns to predict ncRNA genes in *E.coli*'s intergenic regions. The main strengths of the method as compared to other methods are that (i) it uses the DNA sequence directly as input, which helps to reduce any potential bias from input feature selection and encoding (12); (ii) it works well with a much larger number of intergenic sequences (negative examples) than known ncRNA sequences (positive examples) (12); (iii) it is very robust when it comes to noise in the training data, as for instance intergenic regions that actually are ncRNAs; and (iv) it does not rely on sequence conservation to predict ncRNA genes.

The method predicts several hundred intergenic regions to contain ncRNA genes, and over half of these overlap with previous predictions. We test the 10 top-scoring candidates and verify 9 of these by northern analysis. In addition, we test six candidates of varying prediction confidence; three of these are confirmed by northern analysis. Only 6 of these 12 new ncRNA genes have been predicted by previous methods.

Our results indicate that the number of ncRNA genes in *E.coli* is larger than what has previously been estimated (15). This is because the estimates of Zhang and colleagues were partly based on the number of ncRNA genes predicted by more than one method, which, until now, was 95. We have extended

*To whom correspondence should be addressed. Tel: +47 9820 3874; Fax: +47 4559 4458; Email: paal.saetrom@interagon.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

this list by 44%, which is a significant increase. In addition, we have shown that our method detects ncRNA genes that have not been predicted by other methods.

MATERIALS AND METHODS

Sequence data

We downloaded the *E.coli* K-12 genome sequence (16) (U00096.1) and its annotations (release 73) from EMBL's FTP server (http://www.ebi.ac.uk/genomes/bacteria.html). Based on annotations and previous studies (9–11), we collected a set of 154 experimentally verified ncRNA sequences. These sequences consisted of 86 tRNAs, 22 rRNAs and 46 other sRNA genes. Note that one of these sRNAs was the strain-dependent uptR gene (17). The list of ncRNA sequences is given in the Supplementary Material.

Based on the positions of known ncRNA genes and protein coding sequences (CDS), we constructed a set of intergenic sequences (INT) by removing all parts of the genome containing ncRNAs and CDSs, along with 100 nt on each side. This resulted in 942 subsequences totaling 144 520 nt, which increased to 1884 sequences of 289 040 nt when we added the complement of each sequence.

Each ncRNA and INT sequence was then divided into 50 nt sequence windows with 25 nt overlap. If the final window in a sequence had <50 nt, we adjusted the overlap so that the final window also had 50 nt. For example, 90 nt sequences were divided into three 50 nt sequence windows consisting of nucleotides 1–50, 26–75 and 41–90. The 50 nt window size was chosen because the smallest ncRNA in our dataset was 53 nt (dicF). This procedure gave 1795 ncRNA sequence windows and 10 663 INT sequence windows; removing duplicates in the form of identical sequences reduced the number of ncRNA and INT sequence windows to 840 and 10 572. Of the 840 unique ncRNA sequence windows, 53% were from rRNAs, 30% from sRNAs and 17% from tRNAs.

Algorithms

We use a machine learning algorithm called GPboost$_{\mathrm{Reg}}$ to create classifiers that predict whether or not a sequence is an ncRNA gene. The algorithm has previously been used to predict the efficacy of short oligonucleotides in RNAi and antisense experiments (18,19). In the following, we will only give a basic description of the algorithm; interested readers should consult Sætrom (18) and the references therein for a complete description.

GPboost$_{\mathrm{Reg}}$ takes as input a set of positive and negative sequences and creates a classifier that predicts whether or not an unknown sequence belongs to the positive set. Here, the positive and negative sequences are the ncRNA and INT sequence windows described in the previous section. Thus, the classifier created by GPboost$_{\mathrm{Reg}}$ can predict whether or not a given sequence comes from an ncRNA.

To create the classifiers, GPboost$_{\mathrm{Reg}}$ combines genetic programming (GP) (20) and boosting algorithms (21). GP uses simulated evolution in a population of candidate solutions to solve problems, and here, each individual in the population is an expression in a formal query language (whitepaper available on request). GP evaluates how well each candidate solution separates between the positive and negative sequences and uses this fitness information to guide the simulated evolution. That is, our GP solution iteratively (i) selects candidate solutions based on fitness such that more fit solutions have a higher chance of being selected; (ii) introduces random changes in the selected solutions by exchanging subparts of two candidate solutions (crossover) or randomly changing a subpart of a candidate solution (mutation); and (iii) updates the solution population by replacing the old population with the randomly changed candidate solutions. We repeat this process a fixed number of iterations and choose, as the final solution of the GP run, the candidate solution that gave the best performance on the training set.

The classifiers created by our GP algorithm are sequence patterns that can only give binary answers. That is, given a sequence, each pattern answers either 'yes' (1) or 'no' (-1), as to whether the pattern matches parts of the sequence or not. To improve the confidence of our predictions, we combine the GP algorithm with a boosting algorithm. Boosting algorithms join several classifiers into a final weighted average of the individual classifiers such that the performance of the final classifier is increased compared to each of the single classifiers. To do this, the boosting algorithm guides each GP run's search for good solutions by adjusting the relative importance of each sequence in the training set. Then the boosting algorithm assigns a weight to the best expression from the GP run. This weight is based on the expression's performance in the corresponding training set and is assigned such that the output of the final classifier ranges from -1 to 1. As a result, the classifiers created by our algorithm are the weighted average of several different sequence patterns. We will occasionally refer to these classifiers as models. Note that GPboost$_{\mathrm{Reg}}$ uses regularized boosting (22) to handle noise in the training set.

To reduce the time needed to evaluate each individual expression in the GP population, we use a special purpose search processor designed to provide orders of magnitude higher performance than comparable regular expression matchers (23). The increased performance becomes important when the datasets are large, or when many expressions must be evaluated, for instance, in cross-validation experiments or when GP is used as the base learner in a boosting algorithm.

Quality measures

When a model is evaluated on a positive and negative set of sequences, four statistics (counts) can be defined: the number of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*). These represent the positive hits in the positive set, positive hits in the negative set, negative hits in the negative set and negative hits in the positive set, respectively. Several quality measures can be defined from these counts (24). This study uses the Matthews correlation *M* (Equation 1), false positive rate FP_p (Equation 2) and sensitivity *Se* (Equation 3):

$$M = \frac{FP \cdot TN + FP \cdot FN}{\sqrt{(TN + FN) \cdot (TN + FP) \cdot (TP + FN) \cdot (TP + FP)}} \quad \mathbf{1}$$

$$FP_p = \frac{FP}{FP + TN} \quad \mathbf{2}$$

$$Se = \frac{TP}{TP + FN} \quad \mathbf{3}$$

Strain and growth conditions

Escherichia coli K-12 strain MG1655 cells (from overnight cultures were diluted 1/50 in Luria–Bertani (LB) medium and subsequently grown at 37°C) were grown in LB broth and used for inoculation of liquid cultures. Cells were grown in 100-ml batch cultures in 500-ml Erlenmeyer flasks at 37°C with aeration by rotary shaking (250 r.p.m.). The culture media used was LB as described elsewhere (25). Growth was monitored at 600 nm on a Shimadzu UV-1601 UV-visible spectrophotometer. Cells were harvested in four different growth phases: lag ($OD_{600} < 0.2$), log ($0.2 < OD_{600} < 1.0$), early stationary ($1.0 < OD_{600} < 2.0$) and late stationary phase ($OD_{600} > 2.0$).

RNA isolation

Total RNA was isolated from the cells using a procedure based on trizol reagent combined with RNeasy microcolumns (Qiagen). One milliliter of trizol was added per 10^6 cells and stored at room temperature for 5 min; 0.2 μl chloroform was added per ml of trizol and the sample was shaken for 15 s. The sample rested before centrifugation for 15 min at 12000 *g* and 4°C. The aqueous phase was slowly added 1:1 to 70% EtOH to avoid precipitation. The sample was further loaded to the RNeasy column and washed and DNase treated according to the RNeasy protocol (Qiagen). Isolated RNA was resuspended in RNase-free water and quantitated using Eppendorf BioPhotometer.

Oligonucleotides

The complete list of oligonucleotides used to generate probes for northern analysis and primer extension experiments is provided as Supplementary Material.

Northern analysis

RNA samples (~10 μg) were denatured for 10 min at 60°C in a buffer containing 95% formamide, separated on urea–polyacrylamide (8%) gels, and transferred to nylon membranes by electroblotting. Radiolabeled strand-specific RNA probes were synthesized using *in vitro* transcription according to MAXIscript™ (Ambion). Hybridization signals were visualized on Typhoon 9410 (Amersham).

Primer extension assay

Primer extension assay was carried out with AMV reverse transcriptase (Promega), on ~10 μg total RNA and 5′ end-labeled primers. The primers were end-labeled by using [γ^{32}-P]ATP and polynucleotide kinase. Products of the extension reactions were separated on 8% polyacrylamide sequencing gels alongside sequencing reactions performed on the corresponding PCR products from the intergenic regions. Sequencing reactions were carried out with a Thermo Sequenase Radiolabeled Terminator Cycle Sequencing Kit (USB, Amersham).

RESULTS

ncRNA gene predictions

We used a variant of 10-fold cross-validation to train and test our machine learning algorithm (26,27). More specifically, we randomly divided the sets of ncRNA and INT sequence windows into 10 non-overlapping subsets. Then, we iteratively trained classifiers on 8 of the subsets and tested the classifiers on the remaining 2 subsets. We used one of these test subsets to estimate the optimal value of the regularization parameter in the $GPboost_{Reg}$ algorithm and the other test subset as a completely independent test set. We ran this training and testing procedure for 10 iterations such that all the 10 subsets had been used as the independent test set.

To estimate the optimal regularization value, we tried several different values and used the one with the highest average correlation in the 10 'parameter estimation' test subsets. These optimal models had an average correlation of 0.58 on the complete test set, and predicted on average 22 false positive sequence windows in the test subsets. This resulted in an average false positive rate of 2.1%. The models' average sensitivity was 54%. The following sections will examine the predictions in the original ncRNA set, the true positives and false negatives, and the potential new ncRNA genes, the false positives.

The algorithm identifies nearly 80% of the sRNAs in the database. As we used two subsets to test the classifiers, there was some overlap between each of the test sets (each unique sequence was present in two different test sets for two different models). The test set consisted of 840 unique sequences for a total of 1680 sequences: 913 of these were predicted as true positives and 767 were false negatives. When duplicates were removed from these sets, 564 of 840 were positive predictions and 491 of 840 were negative predictions. In other words, 215 sequences were predicted as being both positive and negative. This means that 42% of the sequences were strongly predicted by two models, and 26% were weakly predicted by a single model.

Two of 46 sRNA sequences were completely matched by the models and 10 were completely missed. The complete matches were the partially overlapping rydB and tpe7 found by Wassarman *et al*. (10) and Rivas *et al*. (11), and the misses were micF, oxyS, rybB, ryeE, ryhA, spf, sraB and sraE, and the overlapping ryhB and sraI found by Wassarman *et al*.(10) and Argaman *et al*. (9).

306 potential new ncRNA genes of which 152 confirm previous predictions. The models predicted a total of 438 false positive sequence windows; 57 of these were predicted by two models. Several of the predicted sequence windows overlapped or were located next to each other. When these were joined and treated as one continuous sequence, a total of 306 sequences remained.

A cross-reference of the 306 candidate ncRNA sequences with the list of predicted but unconfirmed ncRNA genes presented in (7) identified that 171 of the sequences overlapped with previous predictions; 152 of these were predicted to be on the same strand. Most of the predictions overlapped with the predictions of Carter and colleagues (12). This was expected, not only because their predictions were the most abundant in our INT set, but also because they base their predictions on the common sequence characteristics of ncRNAs, which is also the essence of our method.

Accounting for the number of predictions made by other methods that were significantly represented (>10 sequences) in our INT set, our predictions support 35, 51, 28 and 41% of the predictions of Rivas *et al*. (11), Carter *et al*. (12),

Chen *et al*. (13) and Tjaden *et al*. (14). Thus, there is relatively good correspondence between our predictions and the predictions of these four methods.

Our results confirm several previous predictions that were not supported by other methods. In total, the intergenic regions in our dataset contained 288 sequences that have been predicted by only one previous method to be part of an ncRNA gene. Our predictions overlapped 123 of these 288 sequences. Excluding the predictions that were unique to the Carter algorithm, our predictions supported 42 of the remaining 166 sequences. Thus, although our predictions increased the list of candidates that are unique to a single study by 15%, we increased the list of candidates predicted by more than one study from 95 to 218 (7). Even when excluding the Carter specific sequences, we increased the list of candidates predicted by more than one study by 44% (7). This is a significant increase.

Table 1 shows the 10 highest scoring intergenic sequence windows (the complete list of predictions are available as Supplementary Material). The table is sorted according to the model output for the highest predicted window in the sequence.

After we started our experiments, several new ncRNA genes in *E.coli* have been identified. Table 2 lists the ncRNA genes that were not included as known ncRNAs in our training set, but that were included with at least 50 nt in our set of intergenic sequences. That is, they were falsely included as negative sequences in the training set. The genes were mainly collected from the *E.coli* genome project's (www.genome.wisc.edu) ASAP database (28) (*E.coli* K-12 Strain MG1655 version m54) and from Refs (7,8).

Although, as Table 2 shows, our method only predicts 2 of the 10 genes to be on the correct strand, the performance is not poorer than that of other methods. For instance, the method of Carter and colleagues (12), which is comparable to our method, predicts only one gene (SroG) correctly. Thus, these genes may be too different to be predictable without combining several of the available methods.

We also cross-referenced our predictions with the unconfirmed transcripts in the cDNA library of Vogel *et al*. (8). Table 3 lists the transcripts that were included with at least 50 nt in our set of intergenic sequences. As the table shows, we predict 5 of the 7 transcripts to be ncRNA genes with the correct orientation. Again, our predictions are comparable to or slightly better than other methods.

Finally, Kawano *et al*. (29) describes several new ncRNA genes. Not all these new ncRNAs were present in our dataset; of the three genes that were present, our predictions match one (RyfB). The other two genes (SokE and SokX), like rdlA, rdlB, rdlC and rdlD, may be involved in anti-sense regulation of hok and ldr (29–31). As these ncRNAs' function is closely linked to their targets' sequences, they may not share many sequence characteristics with other ncRNAs. This can explain why our method has problems predicting these hok/ldr-related ncRNAs.

ncRNA gene validations

To test our predictions, we selected 16 predictions for experimental validation. These included all the top 10 predictions from Table 1 and 6 additional predictions with varying prediction confidence (summarized in Table 4). We chose the 6

Table 1. Top ten predictions sorted by prediction confidence

ID	Position	Length	Strand	Score	Annotation
I001	271879	100	+	0.22	271880–272035 + Carter *et al*.
I002	4230937	150	−	0.22	4230927–4231086 − Carter *et al*.
I003	719883	75	+	0.21	719854–719973 + Carter *et al*.
I004	3766615	50	+	0.21	*Novel*
I005	303544	50	−	0.19	*Novel*
I006	262270	82	−	0.18	*Novel*
I007	4626216	75	+	0.17	*Novel*
I008	1702671	75	+	0.16	1702604–1702818 + Tjaden *et al*.
I009	1859481	125	+	0.16	1859567–1859646 + Carter *et al*.
I010	4527911	50	+	0.15	4527862–4527941 + Carter *et al*.

The given position is the 5′ end for predictions in the positive strand, and the 3′ end for predictions in the negative strand. The score is the classifier output for the highest scoring sequence window in a sequence.

Table 2. Known ncRNA genes included in the set of intergenic sequences

Gene	Overlap	Strand	Prediction	Previous predictions (7)
C0067 (12)	60 of 124	+	Not predicted	n/a
rdlA (30)	66 of 66	+	Predicted 50 nt (−)	?(11), − (12)
rdlB (30)	65 of 65	+	Not predicted	? (11), − (12)
rdlC (30)	67 of 67	+	Not predicted	? (11), − (12)
IS061 (13)	60 of 157	−	Not predicted	n/a
IS092 (13)	116 of 159	−	Not predicted	n/a
rygC (10)	76 of 150	+	Predicted 50 nt (+ and −)	+ (13), − (12)
SroG (8)	110 of 147	−	Predicted 89 nt (−)	− (12)
rdlD (30)	63 of 63	+	Not predicted	− (14), − (12)
SroH (8)	61 of 159	−	Not predicted	+ (13)

The overlap is the number of nucleotides from the ncRNA included as an intergenic sequence. The last column lists the strand and the reference to previous predictions overlapping the gene.

Table 3. Unconfirmed transcripts from (8) included in the set of intergenic sequences

Contig	Overlap	Strand	Prediction	Previous predictions (7)
Contig_440	68 of 105	+	Predicted 50 nt (+) and 50 nt (−)	+ (13), − (12)
Contig_68	76 of 157	+	Predicted 49 nt (+)	+ (14), − (13)
Contig_606	83 of 103	+	Predicted 63 nt (+) and 50 nt (−)	+ (14), − (12)
Contig_223	80 of 141	−	Predicted 50 nt (−)	− (12)
Contig_496	73 of 73	+	Predicted 61 nt (+) and 49 nt (−)	− (14), ± (12)
Contig_286	102 of 102	+	Predicted 50 nt (−)	+ (14)
Contig_181	43 of 43	−	Not predicted	? (11), + (13)

See Table 2 for header explanations.

Table 4. Six predictions with varying confidence experimentally tested in the lab

ID	Position	Length	Strand	Score	Annotation
I014	4373943	60	−	0.14	*Novel*
I016	1218274	50	−	0.14	*Novel*
I035	914278	100	+	0.1	914218–914571 ± Rivas *et al*. 914259–914378 + Carter *et al*.
I044	4366175	50	+	0.1	*Novel*
I209	4006562	50	+	0.025	4006513–4006565 − Carter *et al*.
I211	214141	50	−	0.025	*Novel*

See Table 1 for details on the prediction position.

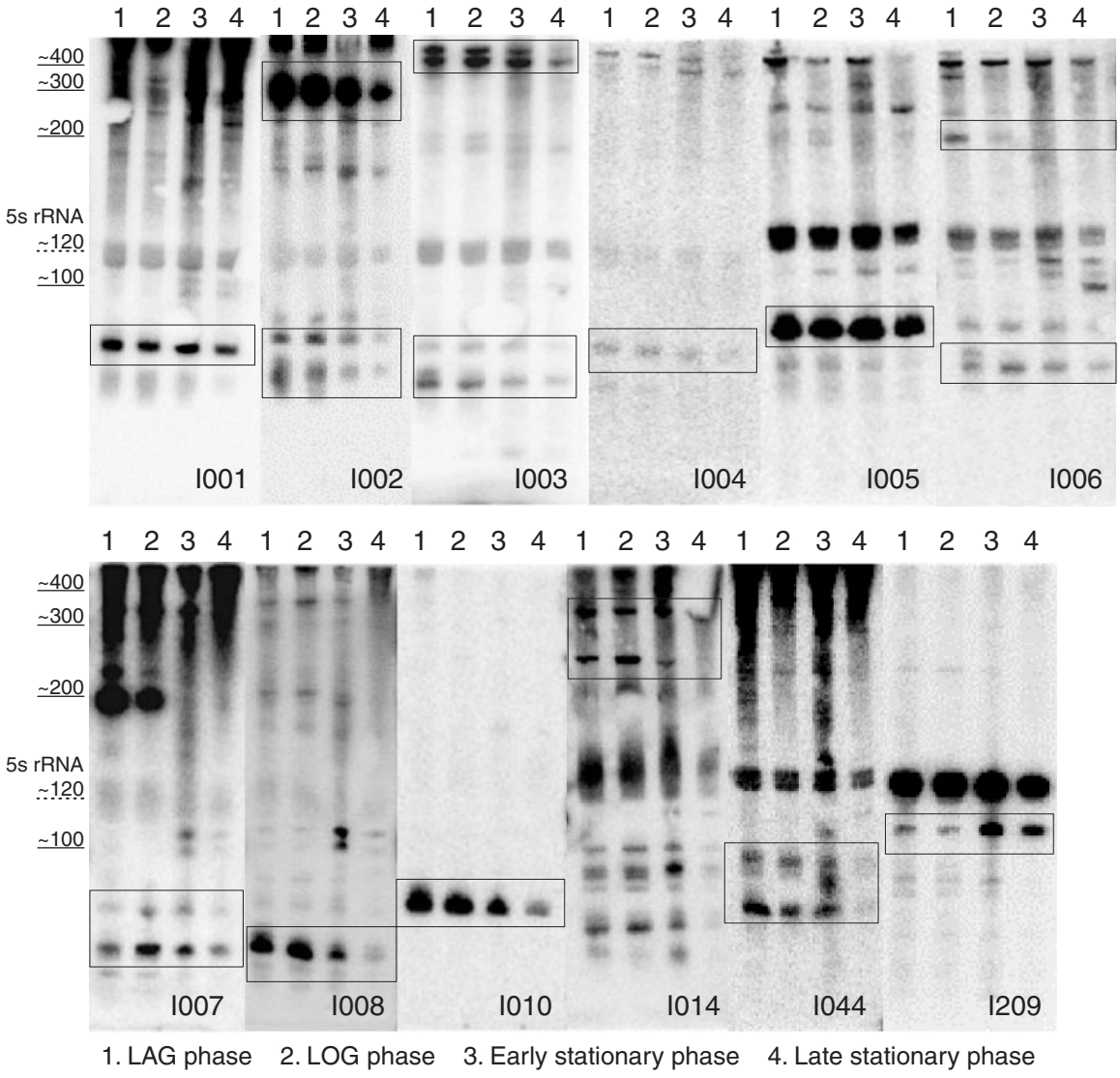

Figure 1. Northern hybridizations of selected predictions against total RNA from lag, log, and early and late stationary phases confirm 12 of 16 selected transcripts. The figure shows the complete northern blots after low stringency wash. The boxed bands indicate the bands that were still present after repeated washes of higher stringency, but the resulting blots are excluded because of poor resolution and picture quality. The indicated sizes are only approximate sizes because these are individual blots lined up together; see Supplementary Figure 2 for size estimates based on each individual blot. Note that most blots have a ~120 nt band that corresponds to 5s rRNA.

additional predictions to have both high and low prediction confidence, and to be a mix of previously predicted and novel candidates. These 6 additions represented a more varying spectrum of predictions than did the top 10 predictions.

Figure 1 shows the results of northern hybridization with strand-specific probes from 12 of the 16 predictions against total RNA from the *E.coli* lag, log, and early and late stationary phases (see Materials and Methods). Most of the 12 confirmed transcripts were differentially expressed in the four phases, which is in agreement with previously known ncRNAs in *E.coli* (8–10). We did not detect transcripts from the four predictions not shown in Figure 1 (data not shown). The absence of detectable transcripts do, however, not imply that the predictions are wrong as some ncRNAs are only expressed under certain conditions [see for example (2,8,10)]. We also tried to map the 5′ start of 4 of the 12 verified transcripts (I001, I002, I004 and I014, chosen because these were a mix of high and low confidence, and previous and novel predictions). We identified potential 5′ start sites for all four transcripts (see Supplementary Material). Based on these results, we estimated the size of three of the transcripts; see Table 5 for additional information.

As Figure 1 shows, we detected more than one band for six of the predictions. These instances of multiple bands were either (i) a large sequence with one or two additional smaller sequences (I002, I003 and I006); (ii) two large sequences (I014); or (iii) two small sequences (I007 and I044). One possible explanation is that the multiple bands are processed or degraded forms of a single transcript. This may be the case for I002 and I014, as we saw only one 5′ start point for each region in the primer extension. These transcripts could be specifically processed by catalytically active enzymes,

Table 5. Transcripts detected by primer extension

Transcript	Strand	5′ start	Predicted distance	Size	5′ gene		3′ gene	
I001	+	271804	75	75	b0257	+	ykfC	+
I002	−	4231116	179	310	b4024 ('lysC')	−	b4025 ('pgi')	+
I004	+	3766359	256	n/a	o153 ('yibG')	+	yibH	−
I014	−	4374139	196	300	o188 ('efp')	+	o155 ('sugE')	+

The table lists the transcripts' 5′ ends; their orientation; the distance between the 5′ ends and the predicted transcripts; the transcripts' estimated size; and the name and orientation of 5′ and 3′ flanking genes (relative to the + strand). Note that the I004 5′ start point overlaps prediction HB_200 of Carter and colleagues (12), but we did not detect any northern signal that corresponded to this 5′ start (see Figure 1).

or unspecifically processed by ribonucleases. Several known ncRNAs in *E.coli* are specifically processed (32), and our results are similar to previously predicted and verified ncRNAs thought to be specifically processed (9).

It is possible that some of the larger transcripts detected could be processed 5′ or 3′ ends of neighboring mRNAs; e.g. I002 overlaps the 5′ CDS of lysC by 6 nt. The neighboring genes that the other large transcripts can and do overlap with (we did not establish the 5′ ends of I003 and I006, but I014 overlaps 4 nt in the 5′ CDS of efp) are on the opposite strand of the verified transcripts. Thus, it is possible that these transcripts can regulate their neighboring genes through an antisense mechanism.

Because the transcripts we have tested have not previously been detected, these transcripts may be unstable or of low abundance and therefore difficult to detect. Such instability may also explain some of the multiple bands. Another possible explanation could be that the strand-specific probes bind to other transcripts, but a Blast (33) search of the probes against the complete *E.coli* genome did not give any matches with *E*-values below 0.1, except for the intended target sites. Thus, it is unlikely that the multiple bands in the northern blots are caused by the probes hybridizing to other complementary transcripts.

Excluding tRNAs and rRNAs improves specificity

Our initial database of ncRNA genes was slightly biased towards rRNA and tRNA genes. As our main focus was to identify other small RNA genes, we did a separate analysis where we trained classifiers exclusively on the sRNA sequences. In this analysis, we used the query language and methodology from Saetrom (18), i.e. a classifier was the average of 10 GPboost runs instead of a single run as in our previous experiments.

Using this approach, we predicted 135 of 255 sRNA sequence windows, which included sequence windows from all but the micF and sraE genes. In addition, the approach identified 140 potential ncRNAs, 69 of which were novel.

A cross-reference of the potential ncRNAs identified by this method with the list of known genes (see Table 2) showed that it had correctly identified the rygC, SroG and rdlD genes. On the other hand, only Contig_496 of the sequences in Table 3 was correctly identified; two other predictions overlapped Contig_440 and Contig_286, but these were on the opposite strand.

As a comparison, we ran an experiment where we again used the approach of Saetrom (18), but also included the tRNAs and rRNAs. We now identified all the ncRNAs in the training set except spf, sraB, sraD and micF, and predicted 401 potential ncRNAs; 168 of these were novel. Although this approach identified slightly fewer of the sRNA genes in the training set compared to the classifiers that were trained only on the sRNA sequences, it identified all the tRNAs and rRNAs; the sRNA-based classifiers only identified 15 of 22 rRNAs and 21 of 86 tRNAs. Thus, as expected, when the rRNAs and tRNAs are excluded from the training set, the resulting classifiers become more specific. In accordance with this, the classifiers trained on the complete ncRNA set identified four of the known ncRNAs in our set of intergenic sequences (rdlA, rygC, SroG and rdlD), and seven of the nine contigs from Table 3 (Contig_440 and Contig_286 were identified on the wrong strand).

DISCUSSION

We have described a novel method for finding non-coding RNA genes and proved its applicability by analyzing *E.coli* intergenic regions, and testing and experimentally confirming 9 of the top 10 scoring predictions and 3 other predictions with lower score. Several groups have searched for new ncRNAs in *E.coli* (8–14), which have resulted in a list of about ∼1000 non-redundant and untested candidates (7). Our predictions mostly confirm the predictions of the other methods, but we also predict several new ncRNA genes, and, as our experimental verifications show, at least six of these new predictions are genuine ncRNAs: 12 of the 16 tested candidates, including 6 novel predictions, were verified. It would therefore be surprising if none of the other candidates are ncRNAs.

Northern analysis and primer extension showed that our method could not completely identify the true transcript of the verified predictions. That is, the algorithm either only predicted a portion of the transcript or misplaced its start and stop site. There are three main reasons for these errors. First, our data set consisted of 50 nt sequence windows with 25 nt overlap. Consequently, we could only predict the correct start and stop site if these regions aligned with any of the sequence windows in our data set. Here, we would expect that only 1 of 25 start sites would align by chance. Second, our algorithm did not recognize all the sequence windows of the known ncRNAs in the training set. We would therefore be surprised if it correctly predicted the complete sequence of any new transcripts. Third, our algorithm is biased in the sense that it will only detect regions that are similar to regions in the known ncRNAs. Thus, the algorithm would have trouble detecting the novel domains in the new transcripts.

Because of these three shortcomings, we did not expect the algorithm to correctly identify the complete sequence of any new transcripts. Rather, we developed the algorithm as

a complementary tool to the existing ncRNA prediction algorithms, which use other features to predict ncRNAs. As an analogy to standard protein coding gene prediction, our algorithm can be considered a content analyzer (34). To get more reliable predictions of complete ncRNAs, we can for example combine our algorithm with algorithms that look for signals such as transcription initiation and termination (9,13). We are currently looking into this.

When comparing our predictions to those of other methods and to the known ncRNAs included in our set of intergenic sequences (see Table 2), we found that some of our predictions were on the opposite strand. In addition, 47 of our predictions overlapped predictions that our algorithm made on the opposite strand (see Supplementary Material). Thus, it appears that the algorithm has problems identifying the correct strand for some transcripts. These results are, however, related to the above discussion on the algorithm's bias: the algorithm will only detect domains that have a similar sequence to those in the known ncRNAs. An ncRNA's function often lies in its secondary structure, however, and in general, several different sequences can fold into the same secondary structure. In particular, for certain sequences both the original and reverse complementary sequence fold into similar secondary structures. Thus, if the reverse complementary of such sequences more closely resembles the known ncRNAs than does the original sequences, our algorithm will predict the reverse complementary sequence to be an ncRNA domain. This is for instance the case for rdlA in Table 2. Our algorithm incorrectly predicted the reverse complementary sequence of rdlA to be an ncRNA, but the secondary structures of the correct sequence mirrors that of the reverse complementary (data not shown).

A recent study uses the sequence conservation of known ncRNA genes and intergenic regions to estimate the number of sRNAs (ncRNAs other than tRNA and rRNA) in *E.coli* to be between 118 and 260 (15). The authors then argue that because the number of sRNA genes that either have been experimentally verified or predicted by at least two different studies in *E.coli* were 150 (at that time), their estimates may be an upper limit to the number of sRNA genes in *E.coli* (15). Following their logic, our results indicate that the number of sRNA genes in *E.coli* may be closer to their highest estimate than to their lowest. This is because we have significantly extended the list of ncRNAs predicted by more than one method, and because we have shown that our method predicts new ncRNAs that have remained undetected by other methods.

To summarize, we have shown that our approach for ncRNA prediction is both accurate and complementary to existing methods. That is, it identifies genuine ncRNA genes, some of which have not been predicted by any other methods.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank K. Lagesen for providing the initial database of *E.coli* ncRNA genes, Y. Esbensen for providing the RNA isolation protocol, and H.E. Krokan and M. Bjørås for valuable comments on the manuscript. The work was supported by the Norwegian Research Council, grants 151899/150, 152020/310 and 152001/150, and the bioinformatics platform at the Norwegian University of Science and Technology, Trondheim, Norway. Funding to pay the Open Access publication charges for this article was provided by the Norwegian Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet*., **2**, 919–929.
2. Wassarman,K.M. (2002) Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell*, **109**, 141–144.
3. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
4. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J., Wheeler,R., Wong,B., Drenkow,J., Yamanaka,M., Patel,S., Brubaker,S., Tammana,H., Helt,G., Struhl,K. and Gingeras,T.R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
5. Mattick,J.S. (2004) RNA regulation: a new genetics? *Nature Rev. Genet*., **5**, 316–323.
6. Wassarman,K.M., Zhang,A. and Storz,G. (1999) Small RNAs in *Escherichia coli*. *Trends Microbiol*., **7**, 37–45.
7. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res*., **31**, 1813–1820.
8. Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G.H. (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res*., **31**, 6435–6443.
9. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G.H., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol*., **11**, 941–950.
10. 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.
11. Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol*., **11**, 1369–1373.
12. Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*., **29**, 3928–3938.
13. Chen,S., Lesnik,E.A., Hall,T.A., Sampath,R., Griffey,R.H., Ecker,D.J. and Blyn,L.B. (2002) A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, **65**, 157–177.
14. Tjaden,B., Saxena,R.M., Stolyar,S., Haynor,D.R., Kolker,E. and Rosenow,C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res*., **30**, 3732–3738.
15. Zhang,Y., Zhang,Z., Ling,L., Shi,B. and Chen,R. (2004) Conservation analysis of small RNA genes in *Escherichia coli*. *Bioinformatics*, **20**, 599–603.
16. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y.S. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
17. Guigueno,A., Dassa,J., Belin,P. and Boquet,P.L. (2001) Oversynthesis of a new *Escherichia coli* small RNA suppresses export toxicity of DsbA'-PhoA unfoldable periplasmic proteins. *J. Bacteriol*., **183**, 1147–1158.
18. Sætrom,P. (2004) Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics*, **20**, 3055–3063.
19. Sætrom,P. and Snøve,Jr,O. (2004) A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun*., **321**, 247–253.
20. Koza,J.R. (1992) *Genetic Programming: On the Programming of Computers by Natural Selection*. MIT Press, Cambridge, MA.

21. Meir,R. and Rätsch,G. (2003) An introduction to boosting and leveraging. In Mendelson,S. and Smola,A. (eds), *Advanced Lectures on Machine Learning*. Springer-Verlag, Vol. 2600, pp. 118–183.
22. Rätsch,G., Onoda,T. and Müller,K.-R. (2001) Soft margins for AdaBoost. *Mach. Learn*., **42**, 287–320.
23. Halaas,A., Svingen,B., Nedland,M., Sætrom,P., Snøve,Jr,O. and Birkeland,O.R. (2004) A recursive MISD architecture for pattern matching. *IEEE Trans. VLSI Syst*., **12**, 727–734.
24. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
25. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
26. Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. [Ser. B] (Methodological)*, **36**, 111–147.
27. Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA, 1137–1143.
28. Glasner,J.D., Liss,P., Plunkett,G.,III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. and Perna,N.T. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res*., **31**, 147–151.
29. Kawano,M., Reynolds,A.A., Miranda-Rios,J. and Storz,G. (2005) Detection of 5′- and 3′-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res*., **33**, 1040–1050.
30. Kawano,M., Oshima,T., Kasai,H. and Mori,H. (2002) Molecular characterization of long direct repeat (LDR) sequences expressing a stable mRNA encoding for a 35-amino-acid cell-killing peptide and a *cis*-encoded small antisense RNA in *Escherichia coli*. *Mol. Microbiol*., **45**, 333.
31. Pedersen,K. and Gerdes,K. (1999) Multiple *hok* genes on the chromosome of *Escherichia coli*. *Mol. Microbiol*., **32**, 1090–1102.
32. Li,Z., Pandit,S. and Deutscher,M.P. (1998) 3′ Exoribonucleolytic trimming is a common feature of the maturation of small, stable RNAs in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*., **95**, 2856–2861.
33. 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.
34. Mathé,C., Sagot,M.-F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.