BIOINFORMATICS
Vol. 00 no. 00 2005
Pages 1–8

PSoL: A Positive Sample Only Learning Algorithm for Finding Non-coding RNA Genes

Chunlin Wang[a], Chris Ding[b], Richard F. Meraz[a], and Stephen R. Holbrook[a]*

[a]Physical Biosciences Division and [b]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

ABSTRACT

Motivation: Small non-coding RNA (ncRNA) genes play important regulatory roles in a variety of cellular processes. However, detection of ncRNA genes is a great challenge to both experimental and computational approaches. In this study, we describe a new approach called positive sample only learning (PSoL) to predict ncRNA genes in the *E. coli* genome. Although PSoL is a machine learning method for classification, it requires no negative training data, which, in general, is hard to define properly and affects the performance of machine learning dramatically. In addition, using the support vector machine (SVM) as the core learning algorithm, PSoL can integrate many different kinds of information to improve the accuracy of prediction. Besides the application of PSoL for predicting ncRNAs, PSoL is applicable to many other bioinformatics problems as well.

Results: The PSoL method is assessed by 5-fold cross-validation experiments which show that PSoL can achieve about 80% accuracy in recovery of known ncRNAs. We compared PSoL predictions with five previously published results. The PSoL method has the highest percentage of predictions overlapping with those from other methods.

Contact: srholbrook@lbl.gov

1 INTRODUCTION

RNA molecules are endowed with extraordinary capacities due to their intrinsic conformational versatility and catalytic abilities. However, their potentials have mostly remained hidden from attention until recently through the discoveries of non-coding RNA (ncRNA) genes. In bacteria, ncRNAs have been found to be involved in the control of transcription (Wassarman and Storz, 2000), RNA processing (Wassarman *et al.*, 1999), RNA stability (Masse and Gottesman, 2002), mRNA translation (Altuvia and Wagner, 2000), and even protein degradation (Gillet and Felden, 2001) and translocation (Keenan *et al.*, 2001). Therefore, ncRNAs play important roles in a variety of cellular processes and correspondingly, efforts to identify the whole set of ncRNAs and then to elucidate their functions are becoming more and more prominent.

However, it is a big challenge to identify the whole set of ncRNA genes in a genome. Most ncRNAs are small and non-susceptible to frame-shift and non-sense mutations, which makes it very difficult to detect using routine biochemical and genetic methods (Hershberg *et al.*, 2003). In addition, ncRNAs have varied stability and are expressed under a variety of environmental and physiological conditions. Therefore, methods such as whole genome microarrays (Tjaden *et al.*, 2002) and the whole genome cloning method (Vogel *et al.*, 2003) are unlikely to fully characterize all ncRNA genes in a genome. The development of computational methods for efficiently finding ncRNA genes in genomic sequences has proven difficult. Unlike protein genes, ncRNA genes lack clear endpoints, vary in size, and have few common statistical features. This poses a great challenge to computational approaches. Despite the difficulties, great efforts have been devoted to predict ncRNA genes by exploring different aspects of properties about known ncRNA genes. Evolutionary conservation of secondary structures provides compelling evidence for biologically relevant RNA function; thus comparative genomics approaches are particularly attractive for ncRNA gene prediction. In a study by Rivas *et al.* (2001), pair stochastic context free grammars were exploited to modeling patterns of co-variation in sequence alignment from related genomes. The program RNAz developed by Washietl *et al.* (2005) basically combines structural conservation and thermodynamical stability of RNA secondary structures in multiple sequence alignments to predict functional RNA structures including ncRNA. Functional sites (i. e. promoter and terminator) are required in ncRNA gene expression. Just as one can reach the melon by following the vine, it is possible to use the predicted signals to approach the boundaries of ncRNA genes. Chen *et al.* (2002) pinpointed ncRNA genes with genomic positions of promoters and terminators, which were predicted based on profile-based methods. The nucleotide composition of known ncRNA genes has been tested to search for discriminative variables between primary sequences of ncRNA genes and intergenic regions in bacterial genome sequence. However, no particular measure stands out to be very discriminative. The combination of some measures such as k-mer (i.e. the usage k nt words) usage might provide a certain level of predictive capability. In addition, different measures often examine different aspects of an actual gene, all of which may complement each other. Therefore, combining different predictive features is highly likely to yield a more accurate prediction. The integrated strategy was initially used to identify ncRNA genes in *E. coli* by Carter *et al.* (2001). Selected discriminative base composition measures and calculated minimum free energies of folding (MFE) were used to train a neural network to distinguish ncRNA from other intergenic sequences. However, less than ten percent of all predictions are shared among different methods above (Hershberg *et al.*, 2003), suggesting that some computational ncRNA gene-finding methods are not highly successful.

We approach the problem of computational prediction of ncRNA genes using a single-class discriminative machine-learning algorithm. Machine-learning involves training a prediction algorithm with knowledge derived from already available data and applying this knowledge to prediction. For this ncRNA prediction problem,

*to whom correspondence should be addressed we try to train a support vector machine (SVM) algorithm (Vapnik, 1995) to distinguish ncRNA genes from intergenic sequences based on statistical differences between biologically relevant, computable representations of these sequences. In general, an SVM is used as a discriminative method to learn a decision boundary from a set of existing examples that can generalize to unseen examples. The performance of an SVM highly depends on the training data set which should consist of examples from all classes to be learned, and have as few misclassifications as possible. However, in many computational biology problems, there are only a limited number of positive (desired) training examples available and the negative examples are difficult to define appropriately.

To overcome the lack of appropriate negative training samples, we developed a new approach called the positive sample only learning (PSoL) algorithm. The PSoL algorithm defines the first set of negative examples by maximizing both the distances between negative sample points to the known positive sample points and the distances among negative samples points simultaneously, and then refines the negative data iteratively using an SVM algorithm based on current positive and negative samples until no additional negative samples can be found according to pre-determined rules. By doing so, a decision boundary is updated iteratively from far away to nearby positive samples, achieving high specificity. In this manuscript we detail the algorithm and apply this approach to the prediction of ncRNA genes in *E. coli*.

2 MATERIALS AND METHODS

2.1 Transformation of biological sequence to feature vectors

The M52 version of the *E. coli* K-12 genome sequence (Blattner *et al.*, 1997) was used to compile a database of ncRNA and non-annotated sequences. The well-characterized ncRNA sequences of *E. coli* were collected based on a literature search (3 rRNA, 20 tRNA and 69 known ncRNA genes, see Supplemental Material 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 functional RNA coding regions from the genome along with a buffer of 50 nucleotides 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. Each RNA or non-annotated intergenic sequence was then divided into sequence windows of 80 nucleotides with a 40-nucleotide overlap between windows (i.e. each window slides 40 nucleotides along the sequence). Any window of < 40 nucleotides was excluded from the study. A total of 5909 windows from each strand (11818 total) were partitioned from the non-coding sequences, while 321 unique RNA sequence windows were generated from the known RNA sequences (after removing redundant RNAs).

Each window was transformed into a feature vector consisting of sequence statistics, the MFE and similarity measurement between related genomes. Sequence statistics were the counts of individual nucleotide (A, C, G, T), dimer (AA, AC $\cdots$ TT) and trimer (AAA, AAC $\cdots$ TTT) in each window. The conservation of the sequence of a window was simply represented by the highest bits score with WU-BLAST $(W = 4)$ between a sequence and the genomic sequence of a reference species. The three reference species are Salmonella typhimurium LT2 (access number NC_003197), *Salmonella typh*i CT18 (access number NC_003198), and *Salmonella typhi* Ty2 (access number NC_004631). The MFE for each window was calculated using the program RNAfold (Washietl *et al.*, 2005) with default parameters. All values were then normalized by dividing by the size of window.

2.2 Feature selection

A total of 88 possible features was generated from the feature extraction method described above. In general, too many features often degrade the performance of the discriminant method by overfitting the training data. Therefore, we picked a small number of features and discard the rest. The most common feature selection involves computing the t-statistic test (for two-class problems) or F-statistic (for multi-class problems) on the class-conditional distributions. Then the features were ranked according to their scores. Those most highly ranked features were then selected.

Both t-statistics and F-statistics assume that for each class, the data follow a normal distribution. In reality, this assumption is not always correct. For this reason, we used a L_1 distance metric between two distributions p, q:

$$d_{L_1}(p,q) = \sum_s |p_s - q_s|.$$

where s is summed over different states. This metric can be viewed as a simplified version of the symmetrized Kullback-Leibler divergence (Kullback and Leibler, 1951): $d_{KL}(p,q) = \sum_s (p_s - q_s) log(p_s/q_s) = \sum_s |p_s - q_s| * |log(p_s/q_s)|$. Since $\log(x)$ is a very slow changing function, we ignore it. The L_1 distance has an intuitive interpretation. If we plot the probability density distribution curves for two different classes, the L_1 distance is the total area sum of the difference between the two curves (see Figs. 1-3). The most discriminant features should have the largest differences on these class-conditional distributions.

The L_1 ranking does not require the underlying data to follow a particular distribution. When the class-conditional distributions are Gaussian, the ranked orders based on t-statistics and on L_1 distance are very similar.

2.3 Learning from partially labeled data

Discriminative machine learning algorithms require labeled data during the training phase. The windows derived from previously identified ncRNA genes were labeled as the positive (+) data. We were trying to distinguish putative ncRNA genes from intergenic sequences. Intergenic sequences contain positive examples (putative ncRNAs) as well as negative examples (sequences that do not encode putative ncRNAs). Therefore, we considered the intergenic sequences to be unlabeled data. Thus our problem became learning from a positively-labeled-only dataset.

2.4 Positive samples only learning

In this problem, we have two types of data: (1) positive data samples and (2) the unlabeled data set, which contains both positives and negatives, and generally much more data than the positive data samples. The goal of PSoL is to predict the positives in the unlabeled data.

PSoL is a challenging problem because there are no negative data. The usual discriminative methods, which require both positive and negative samples for training, cannot be applied to this problem directly. In our earlier approach (Carter *et al.*, 2001), we first took random samples from the unlabeled data and assumed they were negative data. This negative data set plus the true positive data set were used for training the discriminant decision function between the positive and the negative data. This approach is reasonably effective for RNA gene prediction (Carter *et al.*, 2001) since there are many more negatives than positives in the unlabeled sequences.

However, some of the "negative samples" in training the decision function could in fact be positives embedded in the unlabeled data. These wrongly assumed "negative samples" could tilt the decision boundary in an unpredictable way and thus affect the decision boundary significantly.

The key to the success of PSoL is to generate a negative training set without contamination from those ”positives” embedded in the unlabeled data. In this paper, we describe a more sophisticated method to determine the negative training set. The basic spirit of this method has appeared previously (Yu *et al.*, 2002; Yu, 2003; Li and Liu, 2003; Liu *et al.*, 2002).

The method first identifies a small number of data points in the unlabeled data set that are very far away from the positive training data set. In this way, we minimized the possibility of those picked data points to be positive. In addition, we minimized the redundancy in those picked data points by maximizing their mutual distances to achieve a better representativeness for negative data.

Given the small initial negative set, we expanded them in multiple steps, each time picked more data from the currently unlabeled set, using a criteria that they are far-away from the positive training set and close to the current negative set. (The decision function of an SVM gives a convenient measure for the distances to the positives and to the negatives). The negative training set built up in this way will be less contaminated by the positives embedded in the unlabeled dataset.

Once this negative training set is built, we have N: current negative data set, U: remaining unlabeled data set and P: positive data set. The process of predicting positives from the remaining unlabeled dataset is the same as in the two-class prediction.

2.5 Initial negative set selection

2.5.1 Maximum distance minimum redundancy negative set. For the initial negative set, we selected from the unlabeled set m data points that are (1) most dissimilar from the positive set P and (2) least redundant among themselves. We call this maximum distance - minimum redundancy (MDMR) set (Ding and Peng, 2005).

We first defined the distance between a single data point and the positive set, $d(x_i, P)$, as the minimum Euclidean distance between x_i and P:

$$d(x_i, P) = \min_{x_j \in P} \|x_i - x_j\| \tag{1}$$

The maximum distance negative set was constructed by selecting the initial negative set N from the unlabeled set U such that the distance between N and P was maximized:

$$\max_{N \subset U} d(N, P), d(N, P) = \sum_{x \in N} d(x, P) \tag{2}$$

This optimization is trivially solved by picking the N points with largest distance $d(x_i, P)$. However, often the chosen set has many members close to each other and the space represented by N is narrow. From the viewpoint of learning, we may say that there is a certain redundancy in N. To reduce the redundancy, we added a second requirement that maximizes the distance among data points in N:

$$\max_{N \subset U} d(N, N), d(N, N) = \sum_{x_i, x_j \in N} d(x_i, x_j) \tag{3}$$

To satisfy these two criteria simultaneously, we maximize:

$$\max_{N \subset U} [d(N, N) \cdot d(N, P)] \tag{4}$$

The exact solution of Eq.(4), however, is NP hard. We propose the following simple approximate algorithm that is efficient and gives good results in practice.

2.5.2 Forward incremental selection algorithm. The algorithm first selects a point according to Eq.(2). The rest of N is chosen incrementally. Suppose we already have several points in the current negative set N; the new point x_i is selected based on maximum dissimilarity to the positive set:

$$\max_{x_i \in (U-N)} d(x_i, P) \tag{5}$$

And the maximum distance to the current set.

$$\max_{x_i \in (U-N)} \sum_{x_j \in N} d(x_i, x_j) \tag{6}$$

Now Eq.(5) is an exact solution to Eq.(2) and Eq.(6) in an approximate solution to Eq.(3). As in Eq.(4) these two criteria are combined into one:

$$\max_{x_i \in (U-N)} [d(x_i, P) \cdot \sum_{x_j \in N} d(x_i, x_j)] \tag{7}$$

This can be solved by a simple linear search. Once the specified size of N is reached, the algorithm is terminated and we set the initial negative training set $N_{train} = N$.

2.6 Negative set expansion

Given an initial negative set, the PSoL method gradually expands the negative set by classifying more and more unlabeled data points as negative. This is done iteratively using a two-class SVM. At each iteration, an SVM is trained; the decision function values for all remaining unlabeled points are computed, and some of them are classified as negative. Thus $|N|$ is increased and $|U|$ is decreased at each step.

At the stop point, N contains the negative training set and U contains the remaining unlabeled dataset. A final SVM is trained. Based on this, a portion of those in $U - N$ are classified as positive; the remaining ones in $U - N$ are classified as ”undecided”.

2.6.1 Controlled stepwise expansion Given the current negative training set N_{train} and the current unlabeled set U, we perform negative set expansion. We begin by training an SVM on the data $P + N_{train}$ to obtain a large margin decision boundary. The support vectors in N_{train} for this SVM are denoted N_{sv}. All objects in the currently unlabeled set U are tested against the SVM.

We classify unlabeled data points as negative in a conservative and controlled fashion. At an iteration, once the SVM is trained, each unlabeled point will have an decision value $f(x_i)$. Normally, a point x_i is classified as negative if $f(x_i) < 0$. To insure the quality of the negative set, we build a safety margin $h > 0$ by requiring

$$f(x_i) \leq -h, \tag{8}$$

for x_i to belong to the negative set. We typically set $h = 0.2$.

Besides the safety margin, we also control the size of the newly predicted negative samples N_{pred} at each step by setting

$$N_{pred} = \{x_i \mid i \leq r|P| \text{ and } f(x_i) \leq -h\} \tag{9}$$

where r is set to be 3 in most of our experiments.This size control is necessary because the size of unlabeled data samples can be huge compared to that of the positive samples. Therefore the number of newly predicted negative samples is possibly very large in each expansion.

Once N_{pred} are selected, they are added to the current negative set: $N \leftarrow N + N_{pred}$ and they are subtracted from the current unlabeled set: $U \leftarrow U - N_{pred}$

2.6.2 SVM training In SVM training, it is well-known that if the sizes of the classes differ substantively, say $1 : 5$, SVM training typically converges to a solution where all data points in the smaller class are classified as belonging to the larger class.

To overcome this problem, we maintain a current negative training set N_{train} whose size is comparable to $|P|$. At the first iteration, $N_{train} = N$. Later on, after each SVM, the support vectors on the negative side N_{sv} are used to represent the existing negative set N. This is combined with the newly predicted negatives to give the negative training dataset for next round of SVM training: $N_{train} = N_{pred} + N_{sv}$. Since $|N_{pred}| \leq r|P|$, the size of N_{train} is controllable and is maintained in the range where the SVM can be successfully trained with high accuracy.

2.6.3 Stopping criteria of negative expansion Negative set expansion is repeated until the size of the remaining unlabeled set goes below a predefined number, typically about 3 times of the number of expected positives in the unlabeled set. At this last step, the unlabeled data points with the largest positive decision function values are declared as the positives.

2.7 SVM parameter selection

We used the libsvm (Fan *et al.*, 2005) to perform SVM training and predicting. A radial basis function (RBF) kernel was used. There are two parameters for the RBF kernel: γ, which determines the effective range of distances between points, and C, which determines the trade-off between margin maximization and training error minimization. The parameter search is carried out with cross validation. We used a grid-search approach to search for a pair of C and γ with the best performance in cross validation.

It should be emphasized that we fixed parameters for the entire PSoL, i.e., parameters for training the SVM were fixed for each iterative step in the negative set expansion. If we let parameters change during the negative expansion, the data would be overfit and poorer performance in cross validation would result.

3 RESULTS

3.1 Feature selection

For each feature, distributions for positive and unlabeled classes were computed, from which t-score and L_1 score were derived. Detailed distributions for 3 features are shown in Figures 1 - 3. The figures show that distributions follow the normal distributions by varying degrees and the validity of t-score becomes questionable. Since the L_1 measure does not require the underlying data to follow a particular distribution, the L_1 measure can capture the difference. We decided to use 30 features (A, C, G, T, AA, AT, CC, CG, GG, GT, TA, TT, AAA, AAT, ATA, ATT, CCG, CGG, GCC, GGC, GGG, GGT, TAT, TGG, TTA, TTT, MFE, Typhi_CT18, Typhi_Ty2, and Typhi_LT2) with highest L_1 scores.

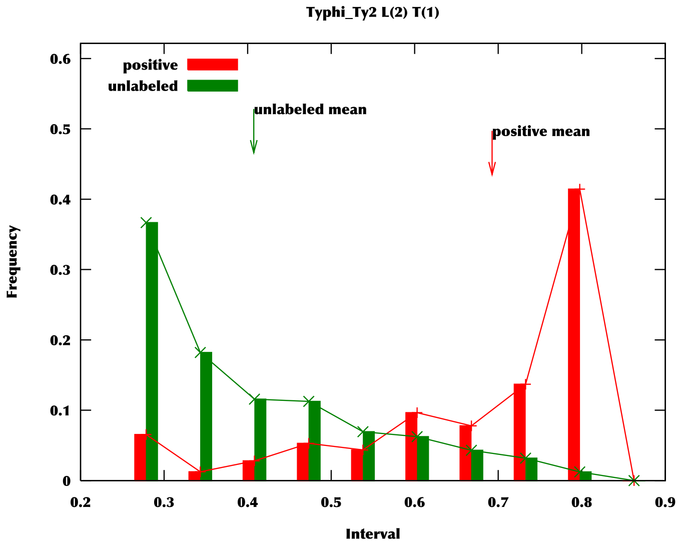

Fig. 1. Distributions (histogram) of normalized bits score for both positive and unlabeled classes on *Salmonella typhi* Ty2 genome sequence (extracted from the best HSP in WU-BLAST search). Both distributions deviate substantially from normal distribution; the sample means shift away from the peak regions. Thus the use of t-score is questionable. L_1 score can capture the difference in distributions.

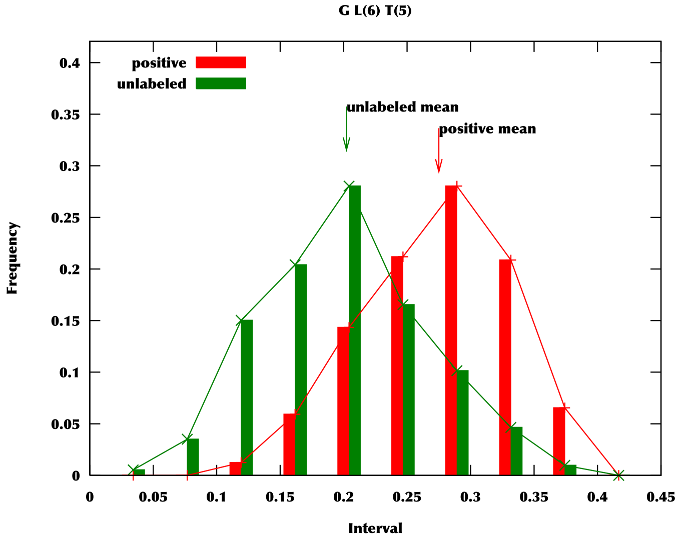

Fig. 2. Distributions of normalized G content for both positive and unlabeled classes. Both t-score and L_1 score can capture the difference in distributions.

3.2 The 5-fold cross validation

In order to calibrate the performance of PSoL on the ncRNA data, we carried out a 5-fold cross validation. Briefly, the positive data were randomly divided into 5 subsets of approximately equal sizes.

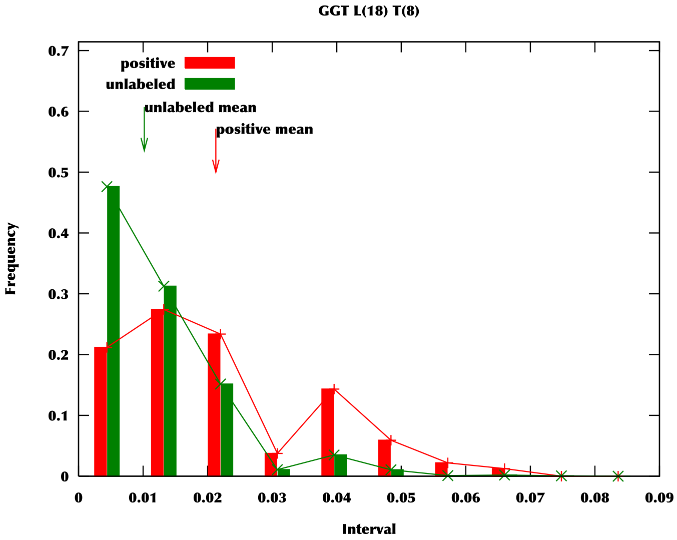

Fig. 3. Distributions of normalized GGT content for both positive and unlabeled classes. t-score can not capture the difference in distributions while L_1 score can.

We ran the validation process 5 times; each time, we merged 4 subsets into positive training data and merged the remaining subset into unlabeled data. We ran the PSoL procedure described above and counted the number of positive samples embedded in the unlabeled data which remain to be ”unlabeled”. Figure 4 shows the results for 5 independent 5-fold cross validation experiments. From those curves, it is apparent that the embedded 64 (321/5) known positives are mostly present in the remaining unlabeled samples as negative expansion proceeds, suggesting that the negative set are not contaminated by the positives. This validates our design of the negative set expansion. When the negative expansion stops at $|U| = 1000$ (1000 samples predicted to be positive), about 80% recovery rate is achieved (see Figure 4). The optimal parameters are $C = 1000$ and $\gamma = 0.04$

ROC curve analysis was carried out to further assess the performance of PSoL. A total of 321 negative control samples were generated by shuffling each positive sample window once using the program SHUFFLE in Sean Eddy’s Squid toolbox (http://hmmer.wustl.edu/) to randomize the sequence while perserving mono- and di-nucleotide composition. The negative samples were marked and put into an unlabeled dataset to do a 5-fold cross validation experiement as described above. The true positive rate and false positive rate were then calculated based on those known positive windows and those negative windows generated by shuffling. The ROC curve of this analysis is shown in Figure 5. When the negative expansion stops at $|U| = 1000$ (1000 samples predicted to be positive), the false positive rate is 6±1 %. Using true positives and true negatives only (igorning the unlabeled category), the average Q^{α} (average of the percentage of correctly predicted positive windows and the percentage of correctly predicted negative windows) (Baldi *et al.*, 2000) of the 5-fold cross validation experiment is 87.3%. We note that this type of estimation of false positive rate can be automatically computed in PSoL and used to help judge when to terminate the process.

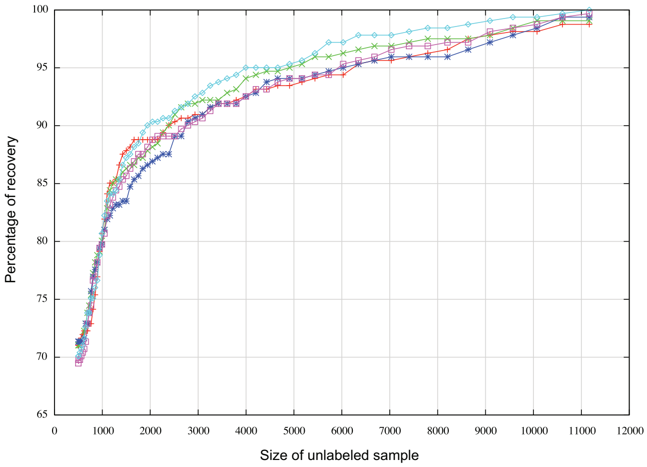

Fig. 4. 5-fold cross validation results. Each curve presents the percentage of correctly recovered positives. We run five 5-fold CV experiments with different random partitions of the positive data. The horizontal coordinate denotes the number of unlabeled samples left after each negative expansion.

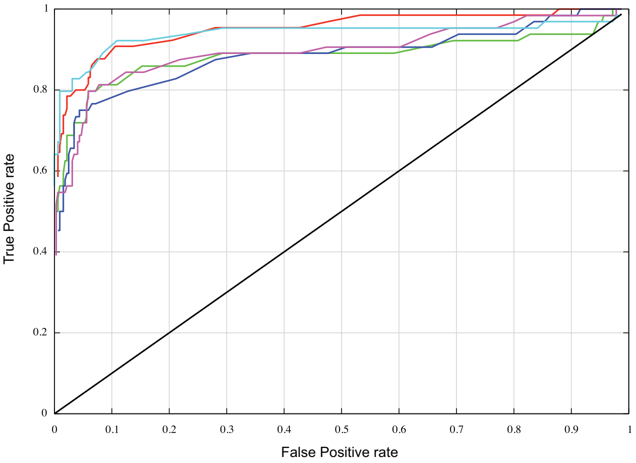

Fig. 5. ROC curves of 5-fold cross validation experiment. Each curve represents the result of one single run of the experiment.

3.3 Prediction

Using the best parameters C and γ from the 5-fold cross-validation experiment, we ran PSoL with all positive data and predicted 1000 windows. This choice is based on the observation that when the number of remaining unlabeled windows is close to 1000, the curves in Fig.4 show a sharp downturn. Since many of these predicted windows were consecutive, we then merged windows that overlapped each other into one. The predicted 1000 windows were assembled into 420 independent RNA sequence segments (details listed in Supplemental Material Table 2).

One of the difficulties in computational prediction of ncRNA genes is the lack of benchmark data to validate the method. Experimental approaches are expensive and time-consuming, therefore only a limited number of predictions were subject to verification

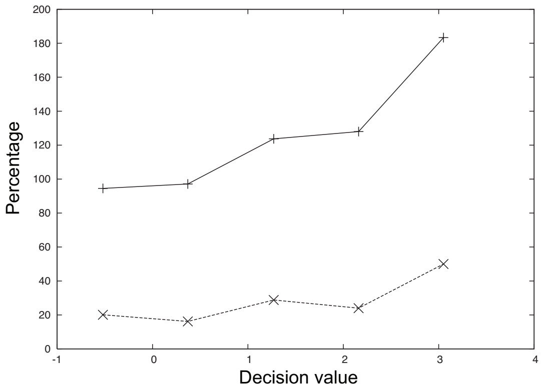

Fig. 6. Percentage of PSoL predictions overlapping with other methods versus decision values. The solid line shows the sum of percentage overlapping with all other 5 methods. The dash line shows the percentage overlapping with Affy result.

using Northern blots or RT-PCR. There are also additional drawbacks to such approaches. Since ncRNA expression may vary according to environmental and physiological conditions, some authentic ncRNA genes might not be detected under experimental conditions. In this study, we compared our predictions to results from previous work. We argued that if our results have more agreements with other studies, that would be a validation of our method. The data used in our comparison were predicted by methods which are listed below.

Abbrev.	Methods	Reference
Affy	microarray experiments	(Tjaden *et al.*, 2002)
QRNA	stochastic context free grammars	(Rivas *et al.*, 2001)
IBIS	promoter and terminator prediction	(Chen *et al.*, 2002)
bGP	boosted genetic programming	(Saetrom *et al.*, 2005)
NNs	neural networks	(Carter *et al.*, 2001)

The results of pairwise comparison are listed in Table 1. Note that a predicted ncRNA gene from one method could overlap with 2 predicted ncRNA genes from another method (see Supplemental Materials figures). In general, PSoL has the largest overlap with other methods. This can be seen clearly from row or column sums. The greater overlap of PSoL with Affy is significant since Affy is the only experimentally based method which results are more reliable.

SVM function values can be used to measure the confidence of prediction. In Fig.6 we show the percentage agreements. It is clear, as the decision values increases, the percentage agreements increases. This suggests that predictions with higher decision values are more likely to be true positives. The rank of each prediction based on its decision value is provided in Supplemental Material Table 2. The secondary structures and their genome schema for the top 5 predictions are also provided in the supplemental data.

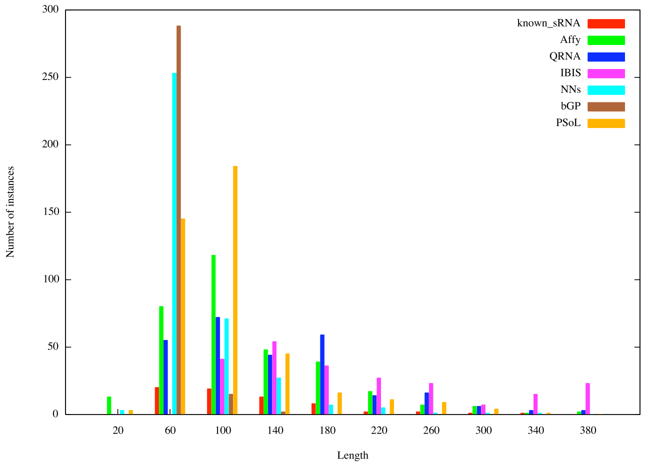

Fig. 7. The length distribution of predicted ncRNA genes and known ncRNA genes (tRNA and rRNA genes are not included).

3.4 Comparison of the statistics of predictions from different methods

Currently, there is no consensus as to the characteristics of an ncRNA gene. In this study, we examined the distribution of length, GC-content, and MFE (minimum free energy) for predictions from different methods mentioned above, as shown in Figures 7, 8, 9. Methods based sequence statistics such as NNs, bGP and PSoL predict more short ncRNA genes. There is less bias in length of prediction from Affy, QRNA and IBIS when compared to the distribution of length of known ncRNA genes. The overall GC content is 50.8% and 40.3% in the *E. coli* K12 genome and in the intergenic regions, respectively. It appears that NNs, bGP and PSoL pick up prediction with slightly higher GC content than the other three methods.

Currently MFE is commonly used as the major predictor for ncRNA gene repdictions. Three out of six methods (PSoL, NNs and bGP) utilize MFE as a prediction parameter. However, as shown in Figure 9, only a small fraction of both known ncRNA genes and predictions from all methods has a very low normalized MFE, suggesting that MFE can not be used as the only predictor of an ncRNA gene.

4 SUMMARY

In summary, the PSoL algorithm addresses two significant concerns in machine learning for biological systems: (1) the uncertainty of the negatives or the lack of negatives, and (2) the overwhelming majority of unlabeled data relative to known positives. This situation is quite common in many bioinformatics problems. We believe our method could provide an effective prediction tool in these difficult situations.

We tested this technique on the prediction of ncRNA genes in the *E.coli* genome sequence solely based on known functional RNA molecules. The 5-fold cross-validation experiments show that PSoL has a recovery rate of 80%. When we compare our predictions with results from previous studies, we find that our prediction has the most overlap with other results, especially with the experimental microarray data, Affy, suggesting the success of this technique.

Table 1. Pairwise overlap between ncRNA prediction methods. We list the number and percentage (in parenthesis) of predictions by the method in the top row overlapping with those by the method in the first column.

	Affy	QRNA	IBIS	NNs	bGP	PSoL	row sum
Affy	-	40 (16.1)	41 (20.1)	69 (19.9)	54 (18.8)	90 (21.4)	294 (96.3)
QRNA	40 (12.8)	-	33 (16.2)	35 (10.1)	23 (8.0)	59 (14.0)	190 (61.1)
IBIS	41 (13.1)	37 (14.9)	-	38 (11.0)	46 (16.0)	48 (11.4)	210 (66.4)
NNs	69 (22.0)	36 (14.5)	40 (19.6)	-	101 (35.2)	149 (35.5)	395 (126.8)
bGP	42 (13.4)	18 (7.3)	37 (18.1)	77 (22.3)	-	90 (21.4)	264 (82.5)
PSoL	92 (29.4)	55 (22.2)	49 (24.0)	149 (43.1)	115 (40.1)	-	460 (158.8)
Column Sum	284 (90.7)	186 (75)	200 (98)	368 (106.4)	339 (118.1)	436 (103.7)	

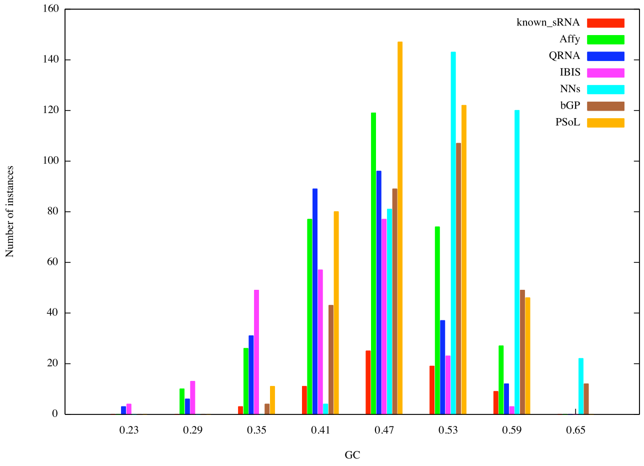

Fig. 8. The GC content distribution of predicted ncRNA genes and known ncRNA genes (tRNA and rRNA genes are not included).

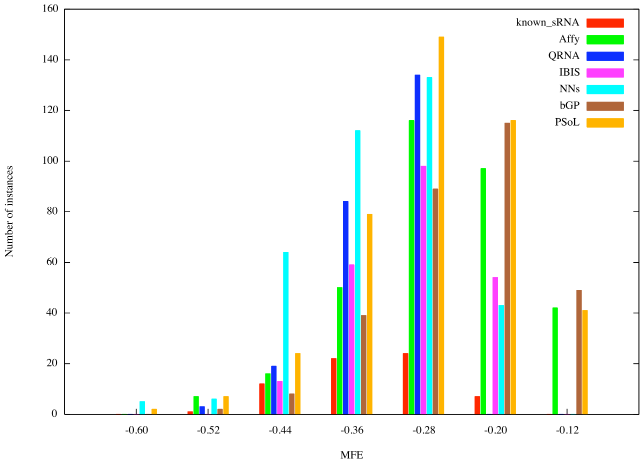

Fig. 9. The normalized (by length) MFE distribution of predicted ncRNA genes and known ncRNA genes (tRNA and rRNA genes are not included).

ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098. We gratefully acknowledge Nancy Schimmelman and Libby Holbrook for editing this manuscript. This work is supported by the National Human Genome Research Institute of the National Institute of Health grant 5RO1H6002665-02.

REFERENCES

Altuvia,S. and Wagner,E.G. (2000) Switching on and off with RNA. *Proc Natl Acad Sci U S A,* **97** (18), 9824–9826.

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.

Blattner,F.R., Plunkett,G.,r., 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. (1997) The complete genome sequence of Escherichia coli K-12. *Science,* **277** (5331), 1453–74.

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** (19), 3928–38.

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** (2-3), 157–77.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol,* **3** (2), 185–205.

Fan,R.E., Chen,P.H. and Lin,C.J. (2005) Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research,* **6**, 1889–1918.

Gillet,R. and Felden,B. (2001) Emerging views on tmRNA-mediated protein tagging and ribosome rescue. *Mol Microbiol,* **42** (4), 879–85.

Gottesman,S. (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet,* **21** (7), 399–404.

Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res,* **31** (7), 1813–20.

Hildebrandt,M. and Nellen,W. (1992) Differential antisense transcription from the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. *Cell,* **69** (1), 197–204.

Keenan,R.J., Freymann,D.M., Stroud,R.M. and Walter,P. (2001) The signal recognition particle. *Annu Rev Biochem,* **70**, 755–75.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86.

Lankenau,S., Corces,V.G. and Lankenau,D.H. (1994) The Drosophila micropia retrotransposon encodes a testis-specific antisense RNA complementary to reverse transcriptase. *Mol Cell Biol,* **14** (3), 1764–75.

Li,X. and Liu,B. (2003) Learning to classify texts using positive and unlabeled data. In *Proc. Joint Conference on Artificial Intelligence (IJCAI-03),* (Gottlob,G. and Walsh,T., eds), pp. 587–594.

Liu,B., Lee,W.S., Yu,P.S. and Li,X. (2002) Partially supervised classification of text documents. In *Proc. 19th Intl. Conf. on Machine Learning* pp. 387–394, Sydney, Australia.

Masse,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A,* **99** (7), 4620–5.

Morfeldt,E., Taylor,D., von Gabain,A. and Arvidson,S. (1995) Activation of alpha-toxin translation in Staphylococcus aureus by the trans-encoded antisense RNA, RNAIII. *Embo J,* **14** (18), 4569–77.

Pfeffer,S., Sewer,A., Lagos-Quintana,M., Sheridan,R., Sander,C., Grasser,F.A., van Dyk,L.F., Ho,C.K., Shuman,S., Chien,M., Russo,J.J., Ju,J., Randall,G., Lindenbach,B.D., Rice,C.M., Simon,V., Ho,D.D., Zavolan,M. and Tuschl,T. (2005) Identification of microRNAs of the herpesvirus family. *Nat Methods,* **2** (4), 269–76.

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** (17), 1369–73.

Saetrom,P., Sneve,R., Kristiansen,K.I., Snove,O.,J., Grunfeld,T., Rognes,T. and Seeberg,E. (2005) Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. *Nucleic Acids Res,* **33** (10), 3263–70.

Sharp,T.V., Schwemmle,M., Jeffrey,I., Laing,K., Mellor,H., Proud,C.G., Hilse,K. and Clemens,M.J. (1993) Comparative analysis of the regulation of the interferon-inducible protein kinase PKR by Epstein-Barr virus RNAs EBER-1 and EBER-2 and adenovirus VAI RNA. *Nucleic Acids Res,* **21** (19), 4483–90.

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** (17), 3732–8.

Vapnik,V.N. (1995) *The nature of statistical learning theory*. SpringerVerlag, New York.

Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res,* **31** (22), 6435–43.

Wagner,E.G. and Simons,R.W. (1994) Antisense RNA control in bacteria, phages, and plasmids. *Annu Rev Microbiol,* **48**, 713–42.

Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A,* **102** (7), 2454–9.

Wassarman,K.M. and Storz,G. (2000) 6S RNA regulates E. coli RNA polymerase activity. *Cell,* **101** (6), 613–23.

Wassarman,K.M., Zhang,A. and Storz,G. (1999) Small RNAs in Escherichia coli. *Trends Microbiol,* **7** (1), 37–45.

Wightman,B., Ha,I. and Ruvkun,G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* **75** (5), 855–62.

Yu,H. (2003) SVMC: single-class classification with support vector machines. In *Proc. of Int. Joint Conf. on Artificial Intelligence*, Acapulco Maxico.

Yu,H., Han,J. and Chang,K.C.C. (2002) PEBL: positive example-based learning for web page classification using SVM. In *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Databases (KDD02)* pp. 239–248 ACM press.

Supplemental Materials for “PSoL: A Positive Sample Only Learning Algorithm for Finding Non-coding RNA Genes”, by C. Wang, C. Ding, R. Meraz, S. Holbrook.

Contains:

- Table 1: Known ncRNA in E.coli used for training;
- Table 2: Predicted ncRNA in E.coli;
- Figure 1: Schema for the top ranked predicted RNA.
- Figure 2: Structure of the top ranked predicted RNA.
- Figure 3: Schema for the 2nd top ranked predicted RNA.
- Figure 4: Structure of the 2nd top ranked predicted RNA.
- Figure 5: Schema for the 3rd top ranked predicted RNA.
- Figure 6: Structure of the 3rd top ranked predicted RNA.
- Figure 7: Schema for the 4th top ranked predicted RNA.
- Figure 8: Structure of the 4th top ranked predicted RNA.
- Figure 9: Schema for the 5th top ranked predicted RNA.
- Figure 10: Structure of the 5th top ranked predicted RNA.

Table 1. Known ncRNA genes in E.coli used for training dataset.

Gene	Start	End	Strand	Type	Reference	Gene	Start	End	Strand	Type	Reference
rrsH	223771	225312	+	rRNA	-	rydB	1762737	1762804	-	sRNA	(Wassarman *et al.*, 2001)
rrlH	225759	228662	+	rRNA	-	rprA	1768396	1768500	+	sRNA	(Majdalani *et al.*, 2001)
rrfH	228756	228875	+	rRNA	-	SroD	1886041	1886126	-	sRNA	(Vogel *et al.*, 2003)
ileV	225381	225457	+	tRNA	-	ryeA	1921090	1921338	+	sRNA	(Wassarman *et al.*, 2001)
alaV	225500	225575	+	tRNA	-	ryeB	1921188	1921308	-	sRNA	(Wassarman *et al.*, 2001)
aspU	228928	229004	+	tRNA	-	IS092	1985862	1986021	-	sRNA	(Chen *et al.*, 2002)
thrW	262095	262170	+	tRNA	-	dsrA	2023249	2023335	-	sRNA	(Sledjeski *et al.*, 1996)
argU	563946	564022	+	tRNA	-	IS102	2069337	2069540	+	sRNA	(Chen *et al.*, 2002)
glnX	695653	695727	-	tRNA	-	ryeC	2151297	2151445	+	sRNA	(Rudd, 1999)
metU	695887	695963	-	tRNA	-	ryeD	2151632	2151774	+	sRNA	(Rudd, 1999)
leuW	696186	696270	-	tRNA	-	ryeE	2165134	2165219	+	sRNA	(Wassarman *et al.*, 2001)
lysT	779777	779852	+	tRNA	-	micF	2311104	2311196	+	sRNA	(Mizuno *et al.*, 1984)
valT	779988	780063	+	tRNA	-	SroE	2638615	2638706	-	sRNA	(Vogel *et al.*, 2003)
serW	925107	925194	-	tRNA	-	ryfA	2651875	2652178	+	sRNA	(Rudd, 1999)
tyrV	1286467	1286551	-	tRNA	-	tke1	2689212	2689360	-	sRNA	(Rivas *et al.*, 2001)
cysT	1989937	1990010	-	tRNA	-	SroF	2689213	2689360	-	sRNA	(Vogel *et al.*, 2003)
glyW	1990065	1990141	-	tRNA	-	ssrA	2753614	2753976	+	sRNA	(Keiler *et al.*, 1996)
asnT	2042571	2042646	+	tRNA	-	sraD	2812822	2812897	+	sRNA	(Argaman *et al.*, 2001)
proL	2284231	2284307	+	tRNA	-	csrB	2922178	2922537	-	sRNA	(Liu *et al.*, 1997)
gltW	2727389	2727464	-	tRNA	-	gcvB	2940718	2940922	+	sRNA	(Urbanowski *et al.*, 2000)
pheV	3108383	3108458	+	tRNA	-	rygA	2974124	2974211	-	sRNA	(Rudd, 1999)
selC	3833849	3833943	+	tRNA	-	rygB	2974332	2974407	-	sRNA	(Rudd, 1999)
trpT	3944581	3944656	+	tRNA	-	ssrS	3054003	3054185	+	sRNA	(Wassarman *et al.*, 2001)
hisR	3980122	3980198	+	tRNA	-	rygC	3054835	3054985	+	sRNA	(Wassarman and Storz, 2000)
sokC	16952	17006	+	sRNA	(Pedersen and Gerdes, 1999)	SroG	3182586	3182734	-	sRNA	(Vogel *et al.*, 2003)
SroA	75516	75608	-	sRNA	(Vogel *et al.*, 2003)	rygD	3192767	3192916	-	sRNA	(Rivas *et al.*, 2001)
t44	189712	189847	+	sRNA	(Rivas *et al.*, 2001)	sraF	3236015	3236203	+	sRNA	(Altuvia *et al.*, 1997)
I006	262270	262352	-	sRNA	(Saetrom *et al.*, 2005)	rnpB	3267857	3268233	-	sRNA	(Brown, 1999)
I001	271879	271979	+	sRNA	(Saetrom *et al.*, 2005)	sraG	3308866	3309039	+	sRNA	(Argaman *et al.*, 2001)
I005	303544	303594	-	sRNA	(Saetrom *et al.*, 2005)	ryhA	3348218	3348325	+	sRNA	(Wassarman *et al.*, 2001)
ffs	475672	475785	+	sRNA	(Brown, 1999)	ryhB	3578554	3578647	-	sRNA	(Wassarman *et al.*, 2001)
SroB	506428	506511	+	sRNA	(Vogel *et al.*, 2003)	IS183	3662494	3662598	+	sRNA	(Chen *et al.*, 2002)
SroC	685904	686066	-	sRNA	(Vogel *et al.*, 2003)	rdlD	3697765	3697828	+	sRNA	(Kawano *et al.*, 2002)
I003	719883	719958	+	sRNA	(Saetrom *et al.*, 2005)	I004	3766615	3766665	+	sRNA	(Saetrom *et al.*, 2005)
rybA	852175	852263	-	sRNA	(Wassarman *et al.*, 2001)	ryiA	3984045	3984216	+	sRNA	(Wassarman *et al.*, 2001)
rybB	887199	887277	-	sRNA	(Wassarman *et al.*, 2001)	I209	4006562	4006612	+	sRNA	(Saetrom *et al.*, 2005)
sraB	1145812	1145980	+	sRNA	(Argaman *et al.*, 2001)	spf	4047479	4047587	+	sRNA	(Mller *et al.*, 2002)
rdlA	1268546	1268612	+	sRNA	(Kawano *et al.*, 2002)	csrC	4048616	4048860	+	sRNA	(Wassarman *et al.*, 2001)
rdlB	1269081	1269146	+	sRNA	(Kawano *et al.*, 2002)	oxyS	4155864	4155973	-	sRNA	(Altuvia *et al.*, 1997)
rdlC	1269616	1269683	+	sRNA	(Kawano *et al.*, 2002)	SroH	4187905	4188065	-	sRNA	(Vogel *et al.*, 2003)
rtT	1286289	1286459	-	sRNA	(Michelsen *et al.*, 1989)	I002	4230937	4231087	-	sRNA	(Saetrom *et al.*, 2005)
IS061	1403676	1403833	-	sRNA	(Chen *et al.*, 2002)	ryjA	4275506	4275645	-	sRNA	(Wassarman *et al.*, 2001)
tke8	1435145	1435252	+	sRNA	(Chen *et al.*, 2002)	I044	4366175	4366225	+	sRNA	(Saetrom *et al.*, 2005)
sokB	1490143	1490195	+	sRNA	(Pedersen and Gerdes, 1999)	I014	4373943	4374003	-	sRNA	(Saetrom *et al.*, 2005)
dicF	1647406	1647458	+	sRNA	(Bouch and Bouch, 1989)	I010	4527911	4527961	+	sRNA	(Saetrom *et al.*, 2005)
I008	1702671	1702746	+	sRNA	(Saetrom *et al.*, 2005)	I007	4626216	4626291	+	sRNA	(Saetrom *et al.*, 2005)

References

Altuvia,S., Weinstein-Fischer,D., Zhang,A., Postow,L. and Storz,G. (1997) A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell,* **90** (1), 43–53.

Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol,* **11** (12), 941–950.

Bouch,F. and Bouch,J.P. (1989) Genetic evidence that DicF, a second division inhibitor encoded by the Escherichia coli dicB operon, is probably RNA. *Mol Microbiol,* **3** (7), 991–994.

Brown,J.W. (1999) The Ribonuclease P Database. *Nucleic Acids Res,* **27** (1), 314.

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** (2-3), 157–77.

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** (2), 333–349.

Keiler,K.C., Waller,P.R. and Sauer,R.T. (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science,* **271** (5251), 990–993.

Liu,M.Y., Gui,G., Wei,B., Preston,J.F., Oakford,L., Yksel,U., Giedroc,D.P. and Romeo,T. (1997) The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in Escherichia coli. *J Biol Chem,* **272** (28), 17502–17510.

Majdalani,N., Chen,S., Murrow,J., John,K.S. and Gottesman,S. (2001) Regulation of RpoS by a novel small RNA: the characterization of RprA. *Mol Microbiol,* **39** (5), 1382–1394.

Michelsen,U., Bsl,M., Dingermann,T. and Kersten,H. (1989) The tyrT locus of Escherichia coli exhibits a regulatory function for glycine metabolism. *J Bacteriol,* **171** (11), 5987–5994.

Mizuno,T., Chou,M.Y. and Inouye,M. (1984) A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci U S A,* **81** (7), 1966–1970.

Mller,T., Franch,T., Udesen,C., Gerdes,K. and Valentin-Hansen,P. (2002) Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev,* **16** (13), 1696–1706.

Pedersen,K. and Gerdes,K. (1999) Multiple hok genes on the chromosome of Escherichia coli. *Mol Microbiol,* **32** (5), 1090–1102.

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** (17), 1369–73.

Rudd,K.E. (1999) Novel intergenic repeats of Escherichia coli K-12. *Res Microbiol,* **150** (9-10), 653–664.

Saetrom,P., Sneve,R., Kristiansen,K.I., Snove,O.,J., Grunfeld,T., Rognes,T. and Seeberg,E. (2005) Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. *Nucleic Acids Res,* **33** (10), 3263–70.

Sledjeski,D.D., Gupta,A. and Gottesman,S. (1996) The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in Escherichia coli. *EMBO J,* **15** (15), 3993–4000.

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** (4), 856–868.

Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res,* **31** (22), 6435–43.

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** (13), 1637–51.

Wassarman,K.M. and Storz,G. (2000) 6S RNA regulates E. coli RNA polymerase activity. *Cell,* **101** (6), 613–23.

Table 2. Predicted sRNA genes by PSoL, ranked according to decision values.

Start	End	Len	Rank	Start	End	Len	Rank	Start	End	Len	Rank	Start	End	Len	Rank	Start	End	Len	Rank
11996	12113	117	221	1061671	1061723	52	360	2087280	2087360	80	377	3096472	3096527	55	316	3920130	3920210	80	162
57159	57279	120	58	1067501	1067581	80	282	2088118	2088164	46	250	3100920	3100981	61	204	3923311	3923590	279	203
59369	59449	80	299	1078395	1078475	80	123	2099580	2099660	80	348	3136587	3136644	57	131	3930552	3930792	240	47
63314	63379	65	179	1108414	1108494	80	410	2116596	2116652	56	237	3147576	3147627	51	375	3938856	3938936	80	264
77389	77549	160	277	1112679	1112752	73	314	2135395	2135475	80	122	3151488	3151528	40	86	3938976	3939056	80	111
89282	89362	80	283	1156967	1157042	75	315	2137557	2137677	120	246	3154582	3154704	122	118	3939096	3939381	285	7
89522	89584	62	201	1160984	1161058	74	406	2163593	2163640	47	320	3175976	3176075	99	95	3941222	3941276	54	175
113349	113394	45	371	1169647	1169691	44	290	2183371	2183451	80	12	3181429	3181509	80	402	3958167	3958242	75	202
127757	127862	105	55	1184867	1184947	80	189	2209134	2209183	49	417	3184111	3184153	42	313	3975411	3975491	80	293
131470	131565	95	6	1195726	1195806	80	234	2212794	2212836	42	401	3198696	3198776	80	169	3988450	3988530	80	249
164624	164680	56	190	1195926	1196006	80	276	2213667	2213715	48	157	3208274	3208372	98	185	3992118	3992322	204	96
190649	190807	158	139	1210652	1210772	120	147	2227349	2227408	59	216	3215093	3215147	54	376	3998758	3998988	230	10
191758	191805	47	64	1211636	1211716	80	354	2230798	2230848	50	307	3245118	3245198	80	363	4005351	4005551	200	209
192670	192750	80	78	1212460	1212501	41	136	2234650	2234713	63	391	3267518	3267807	289	130	4006662	4006726	64	346
194767	194847	80	399	1222681	1222761	80	291	2246602	2246682	80	403	3279316	3279396	80	311	4008483	4008563	80	252
222695	222783	88	171	1225713	1225773	60	387	2261565	2261645	80	164	3310869	3310933	64	238	4010496	4010593	97	65
223458	223721	263	9	1236674	1236744	70	394	2278553	2278602	49	74	3315703	3315804	101	75	4014829	4014870	41	82
225625	225709	84	228	1260049	1260101	52	79	2302583	2302663	80	419	3319562	3319642	80	247	4025108	4025149	41	144
239134	239214	80	366	1261148	1261199	51	76	2302703	2302783	80	322	3320196	3320316	120	63	4032083	4032147	64	267
243429	243493	64	344	1278621	1278701	80	198	2302943	2303063	120	407	3325774	3325830	56	305	4032792	4033070	278	33
245855	245911	56	259	1285839	1285882	43	304	2317898	2318013	115	167	3326431	3326554	123	225	4034974	4035049	75	398
253251	253331	80	331	1286601	1286711	110	66	2378488	2378608	120	212	3352095	3352217	122	85	4038266	4038426	160	22
253371	253417	46	50	1294551	1294619	68	180	2378648	2378692	44	115	3375185	3375265	80	188	4049486	4049569	83	26
263321	263430	109	390	1297514	1297594	80	385	2389395	2389482	87	28	3376336	3376455	119	336	4051248	4051399	151	156
271529	271609	80	245	1306799	1306879	80	177	2391111	2391175	64	364	3377682	3377770	88	133	4054049	4054129	80	286
271649	271769	120	215	1308503	1308543	40	200	2403382	2403462	80	230	3382193	3382288	95	149	4055704	4055937	233	49
274431	274475	44	285	1328822	1329022	200	98	2404751	2404831	80	409	3389993	3390044	51	416	4076508	4076588	80	196
279349	279549	200	101	1333183	1333263	80	48	2405471	2405531	60	365	3398914	3398979	65	92	4083452	4083546	94	41
288436	288475	39	223	1333623	1333743	120	242	2411202	2411282	80	126	3402504	3402584	80	326	4095079	4095159	80	217
289619	289699	80	163	1349313	1349381	68	372	2428833	2428913	80	287	3420880	3421009	129	5	4103301	4103348	47	324
312611	312691	80	260	1355224	1355304	80	369	2438270	2438355	85	243	3424470	3424545	75	412	4105001	4105081	80	194
334336	334416	80	420	1360542	1360662	120	112	2454882	2454962	80	87	3426449	3426607	158	67	4115920	4116000	80	368
440657	440723	66	89	1384646	1384694	48	338	2459188	2459270	82	32	3440158	3440205	47	350	4126025	4126105	80	332
454063	454143	80	205	1407107	1407267	160	30	2463077	2463157	80	113	3450957	3451095	138	2	4129937	4130017	80	229
454223	454307	84	61	1416475	1416522	47	273	2494755	2494835	80	153	3476224	3476304	80	174	4130097	4130146	49	99
455745	455851	106	284	1439018	1439178	160	84	2494955	2495027	72	257	3476344	3476389	45	218	4135327	4135407	80	279
460516	460625	109	300	1439218	1439295	77	233	2519438	2519558	120	165	3483505	3483625	120	35	4160899	4161168	269	56
460997	461089	92	120	1485157	1485209	52	261	2526014	2526131	117	83	3483665	3483707	42	39	4163914	4164188	274	43
480382	480428	46	329	1486149	1486206	57	274	2531450	2531530	80	44	3494644	3494842	198	46	4166116	4166170	54	178
496308	496349	41	114	1577536	1577607	71	359	2539323	2539649	326	13	3516953	3517053	100	105	4169385	4169545	160	23
547621	547788	167	140	1617062	1617142	80	155	2541678	2541798	120	207	3523917	3523997	80	384	4174757	4174887	130	81
563753	563833	80	374	1620713	1620934	221	280	2559048	2559128	80	309	3537394	3537474	80	341	4177251	4177491	240	1
573652	573732	80	355	1634721	1634801	80	248	2588778	2588898	120	121	3550156	3550236	80	192	4178553	4178773	220	90
579639	579719	80	227	1642577	1642625	48	166	2590844	2591042	198	102	3571185	3571305	120	94	4193727	4193860	133	34
585266	585320	54	31	1643016	1643093	77	251	2613951	2614031	80	388	3572561	3572641	80	356	4197722	4197809	87	53
596246	596304	58	21	1647115	1647195	80	292	2642353	2642403	50	226	3576487	3576531	44	306	4205160	4205240	80	210
608570	608632	62	333	1647315	1647356	41	418	2651689	2651809	120	100	3598464	3598544	80	318	4205360	4205675	315	27
613252	613330	78	265	1671735	1671815	80	183	2660201	2660281	80	176	3635170	3635222	52	370	4207517	4207572	55	152
631512	631562	50	54	1694385	1694436	51	235	2687576	2687641	65	117	3637640	3637691	51	288	4216082	4216125	43	214
638781	638896	115	18	1710632	1710712	80	325	2693853	2693909	56	268	3645491	3645571	80	297	4237519	4237599	80	392
643240	643320	80	24	1735444	1735524	80	302	2698073	2698153	80	182	3645931	3646011	80	408	4237679	4237759	80	353
663236	663275	39	88	1735684	1735818	134	107	2702213	2702293	80	342	3661207	3661327	120	289	4242704	4242758	54	382
668201	668281	80	68	1739276	1739387	111	269	2708322	2708390	68	191	3667108	3667172	64	361	4244048	4244128	80	281
694228	694274	46	240	1753455	1753535	80	184	2729228	2729508	280	57	3673828	3673870	42	327	4244208	4244313	105	108
696406	696486	80	224	1753575	1753672	97	106	2751527	2751765	238	8	3679660	3679741	81	91	4266896	4266943	47	197
710738	710778	40	109	1766799	1766919	120	40	2763344	2763424	80	381	3681127	3681207	80	414	4291968	4292010	42	103
712075	712155	80	219	1770359	1770439	80	367	2773092	2773172	80	239	3693864	3693944	80	379	4327906	4327986	80	17
728005	728085	80	73	1777375	1777455	80	386	2773692	2773772	80	310	3705464	3705624	160	146	4329632	4329709	77	125
753741	753861	120	135	1786352	1786409	57	72	2775973	2776117	144	272	3705944	3706195	251	69	4339336	4339439	103	206
754101	754350	249	19	1797016	1797136	120	253	2781318	2781398	80	158	3713959	3714039	80	393	4346372	4346452	80	405
773875	773925	50	357	1819693	1819773	80	295	2781478	2781558	80	70	3717487	3717607	120	77	4349330	4349371	41	337
780191	780241	50	36	1846750	1846811	61	298	2815655	2815756	101	29	3719728	3719808	80	312	4368160	4368216	56	42
780495	780542	47	59	1859566	1859646	80	335	2815932	2816031	99	38	3723077	3723197	120	128	4372022	4372102	80	415
780965	781085	120	186	1860583	1860745	162	60	2816346	2816445	99	52	3734856	3734936	80	271	4403653	4403718	65	193
802593	802652	59	16	1864826	1864882	56	220	2866826	2866866	40	395	3735016	3735076	60	334	4422579	4422646	67	159
812300	812380	80	168	1876934	1876981	47	380	2870893	2870973	80	400	3748661	3748708	47	317	4436195	4436235	40	323
816120	816217	97	45	1891905	1891985	80	173	2875731	2875891	160	351	3752508	3752553	45	321	4460404	4460484	80	345
819861	819941	80	296	1894862	1894906	44	236	2876011	2876091	80	397	3769781	3769858	77	383	4464990	4465149	159	199
836789	836838	49	396	1899699	1899819	120	161	2902046	2902126	80	330	3771911	3772001	90	340	4481898	4481958	60	195
837278	837363	85	328	1899979	1900022	43	301	2902246	2902406	160	187	3774704	3774784	80	378	4483569	4483649	80	241
921863	921943	80	160	1903453	1903573	120	124	2907778	2907866	88	127	3774904	3774976	72	413	4506275	4506395	120	294
925756	925836	80	389	1923043	1923082	39	254	2920292	2920412	120	148	3782119	3782161	42	349	4508151	4508208	57	373
931603	931768	165	20	1927941	1928008	67	150	2920452	2920507	55	255	3782692	3782772	80	308	4516500	4516580	80	213
962941	963001	60	14	1932758	1932813	55	137	2941260	2941309	49	172	3805775	3805855	80	270	4531442	4531802	360	51
982167	982248	81	62	1948596	1948676	80	142	2962289	2962333	44	278	3809351	3809468	117	15	4532282	4532362	80	244
983650	983692	42	263	1957034	1957154	120	110	2964109	2964160	51	145	3834033	3834113	80	143	4549115	4549155	40	154
986295	986375	80	339	1980461	1980528	67	352	2967129	2967209	80	151	3850745	3850865	120	80	4580668	4580769	101	37
988258	988327	69	362	1990191	1990242	51	116	2967409	2967569	160	141	3850985	3851345	360	4	4588914	4588994	80	222
989629	989709	80	347	2011126	2011201	75	256	3010544	3010585	41	208	3865099	3865179	80	258	4603362	4603442	80	343
1006913	1007017	104	97	2050206	2050248	42	134	3030925	3031035	110	232	3881407	3881607	200	119	4604100	4604188	88	411
1019326	1019446	120	104	2065953	2066033	80	319	3053520	3053582	62	3	3881727	3881887	160	211	4608862	4608915	53	71
1019526	1019583	57	262	2066393	2066513	120	132	3069314	3069394	80	25	3882531	3882655	124	11	4609700	4609900	200	181
1031265	1031312	47	231	2076526	2076606	80	266	3071761	3071921	160	129	3906044	3906127	83	275	4628141	4628221	80	404
1050946	1051020	74	303	2085140	2085260	120	170	3079704	3079756	52	93	3911346	3911408	62	138	4630705	4630752	47	358

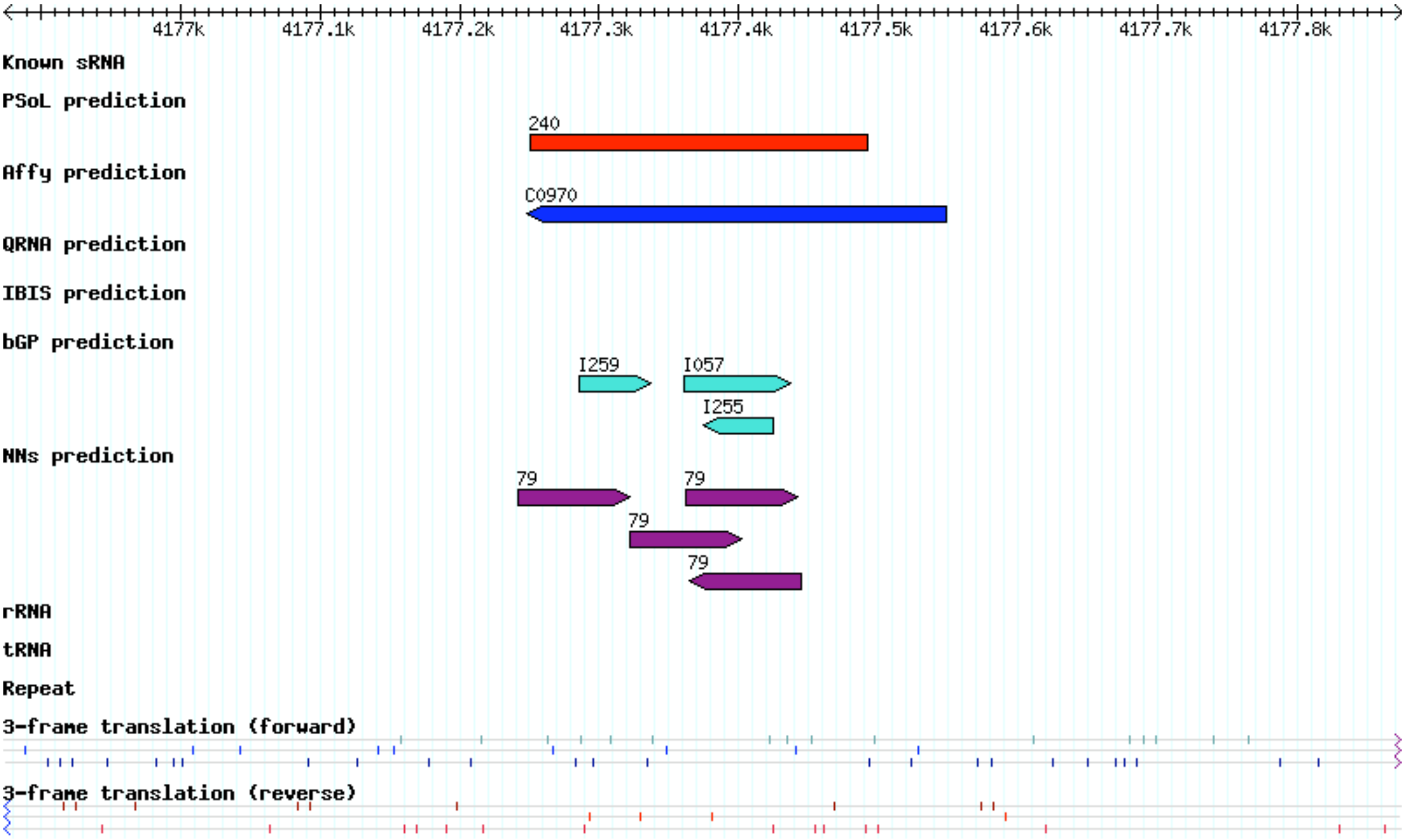

Figure 1: The schema of the first ranked prediction. The methods are indicated on the left side.

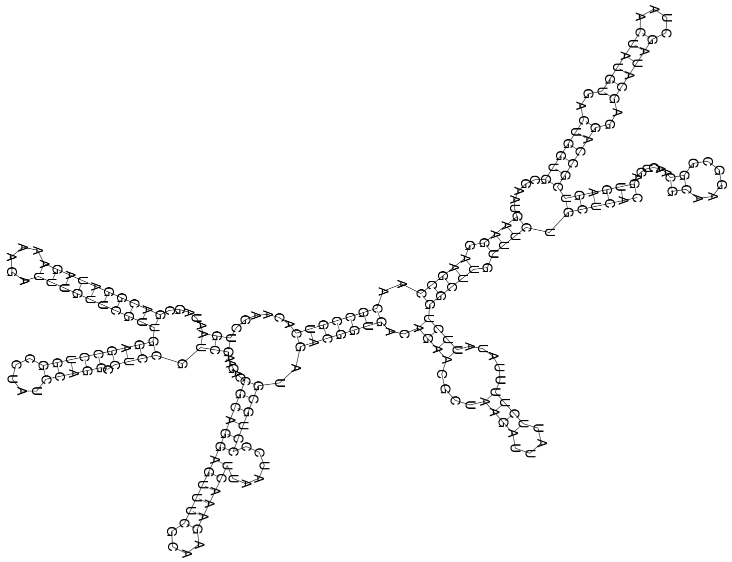

Figure 2: The structure of the first ranked prediction (based on the sequence on the forward strand)

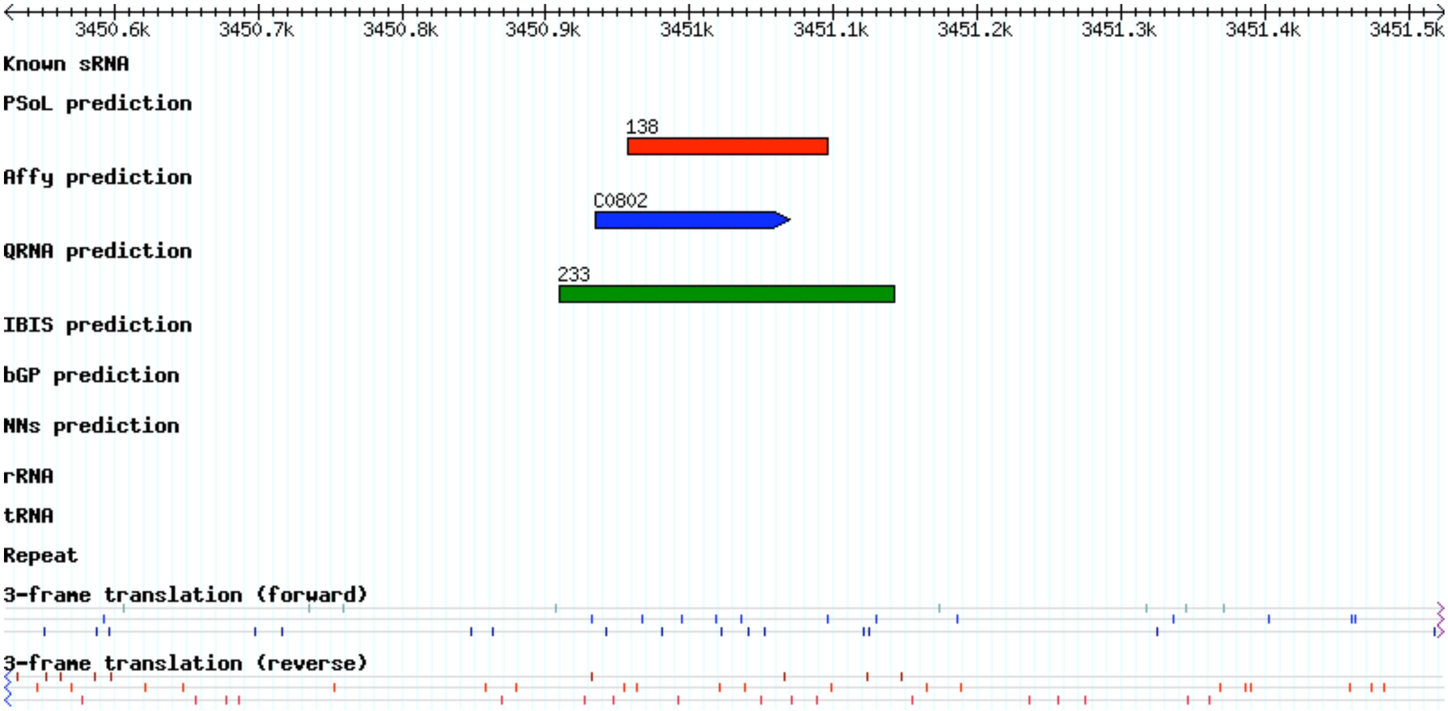

Figure 3: The schema of the second ranked prediction

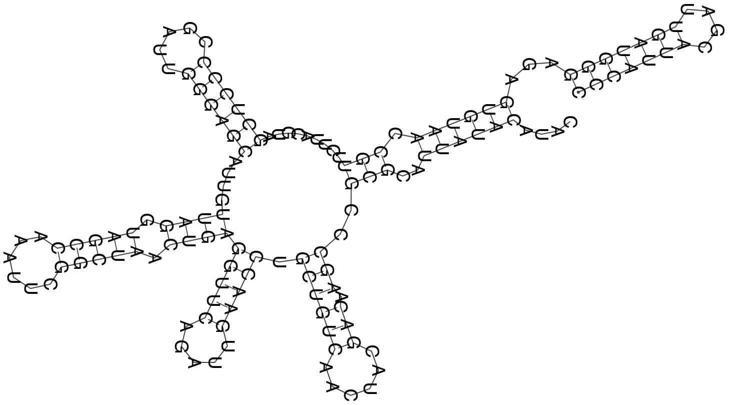

Figure 4: The structure of the second ranked prediction (based on the sequence on the forward strand).

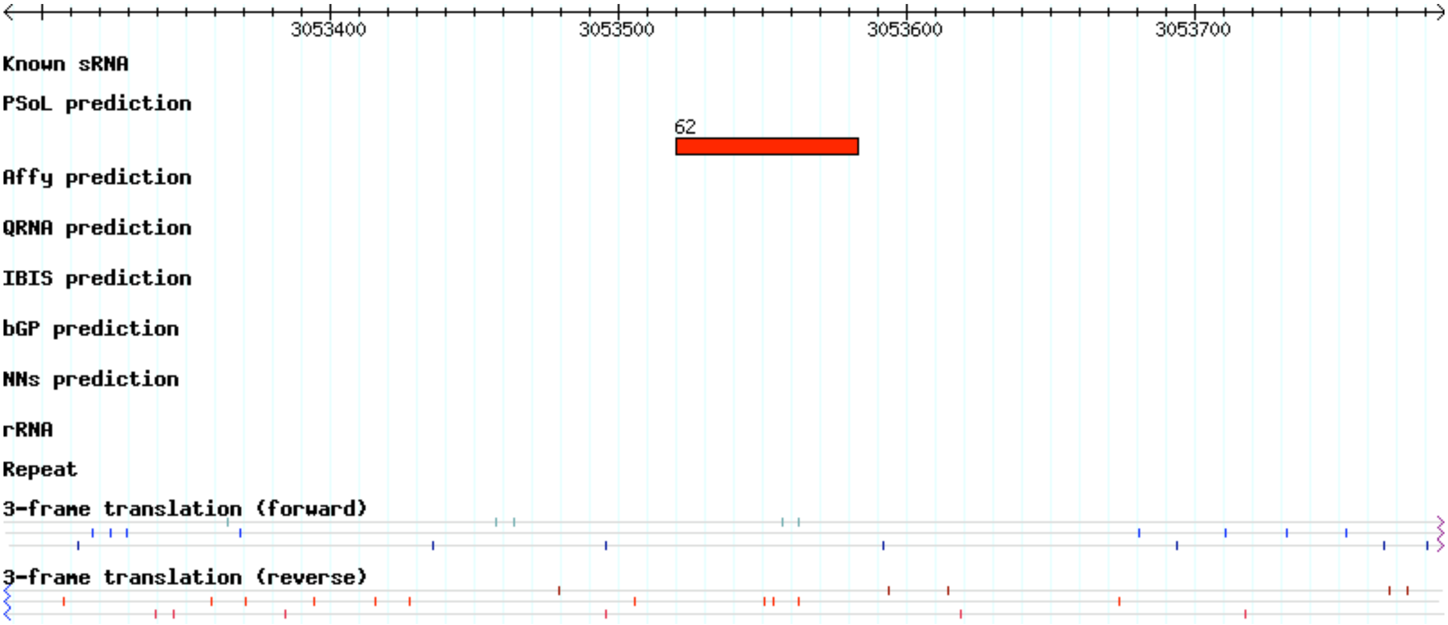

Figure 5: The schema of the third ranked prediction

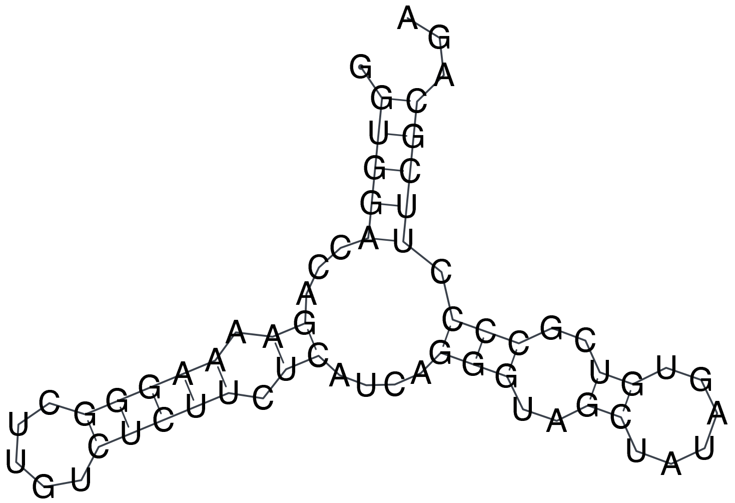

Figure 6: The structure of the third ranked prediction (based on the sequence on the forward strand).

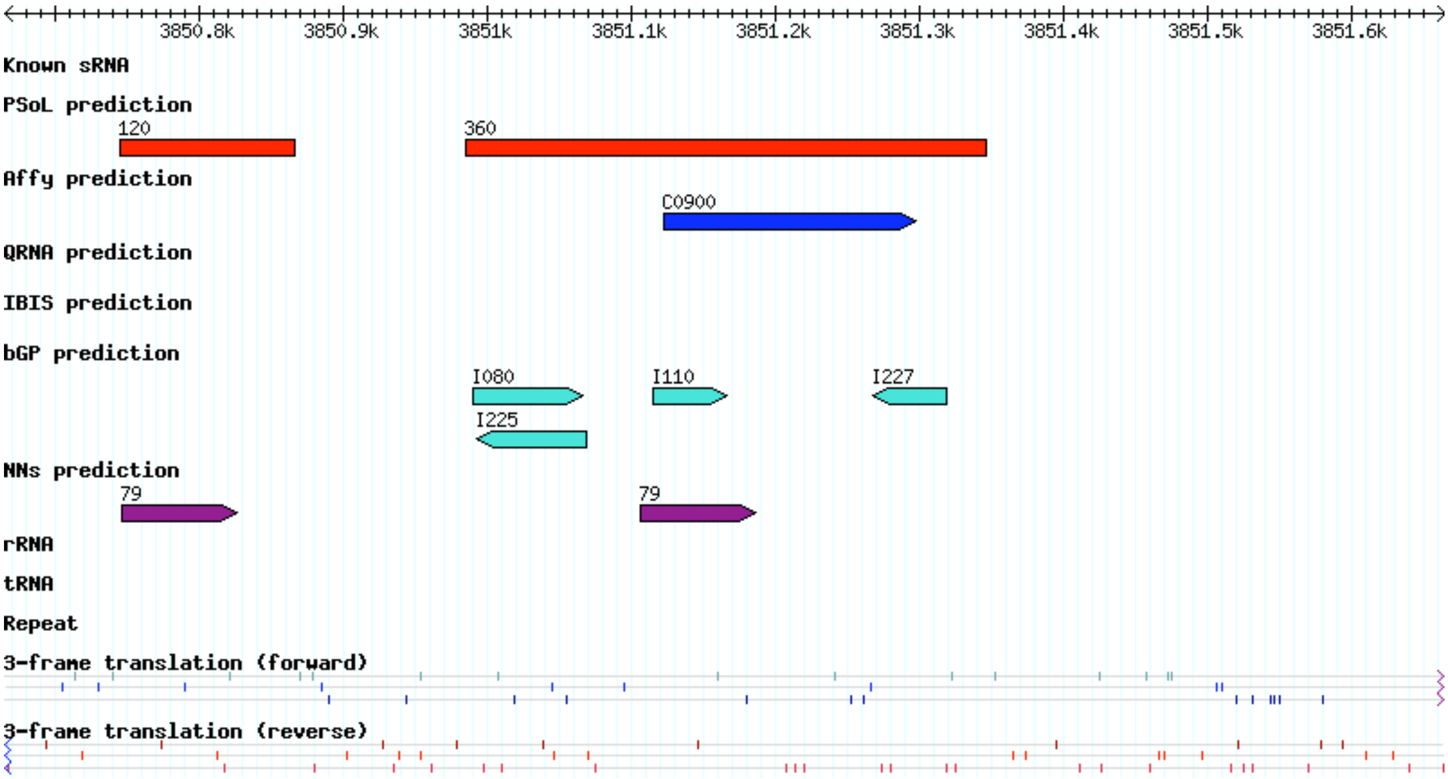

Figure 7: The schema of the fourth ranked prediction

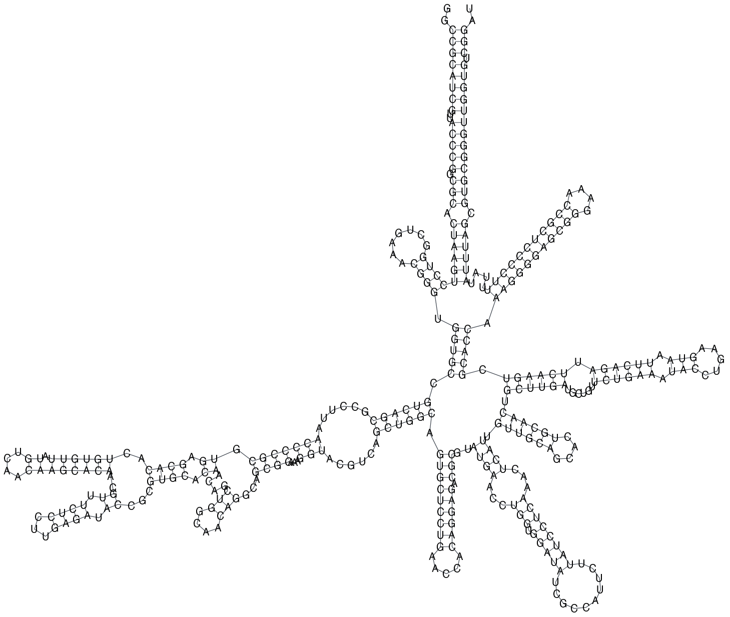

Figure 8: The structure of the fourth ranked prediction (based on the sequence on the forward strand).

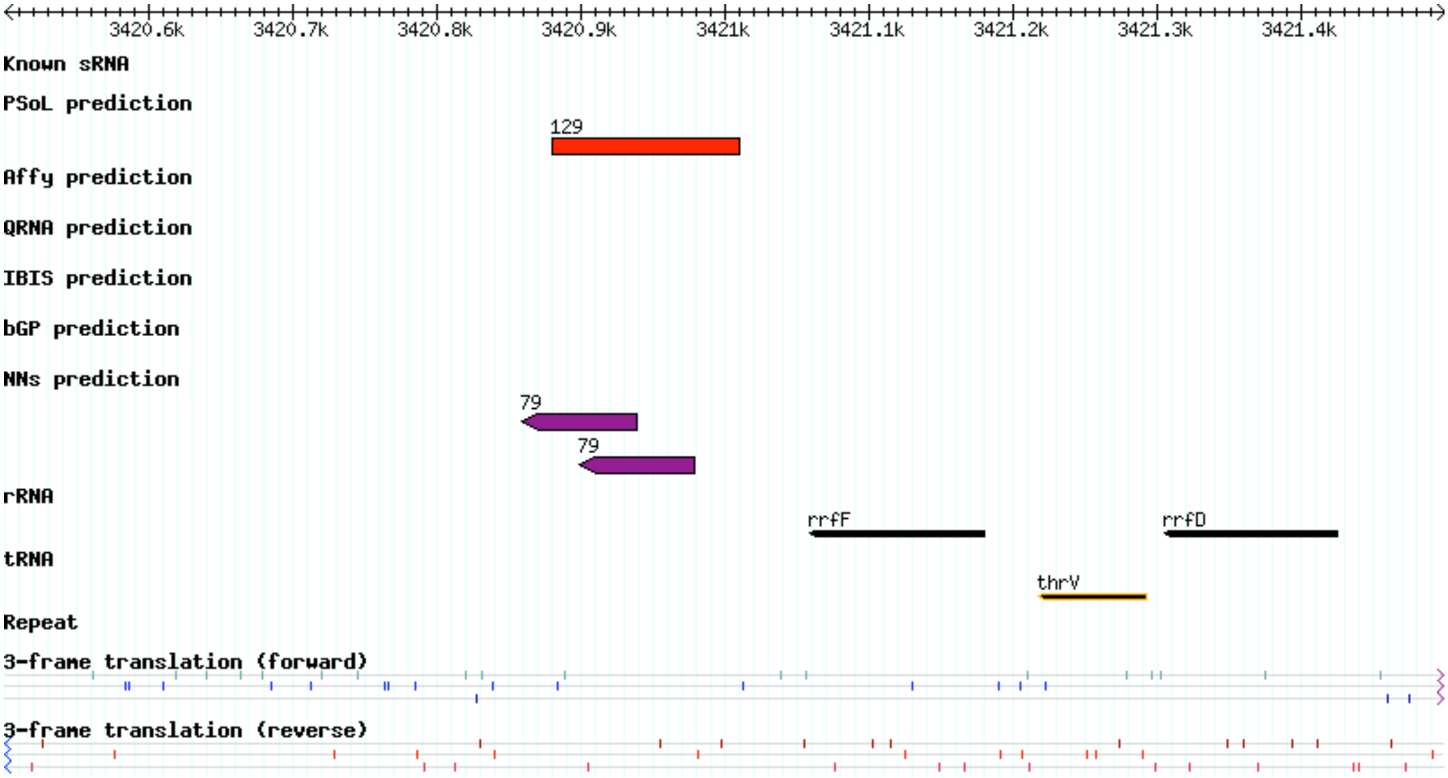

Figure 9: The schema of the fifth ranked prediction

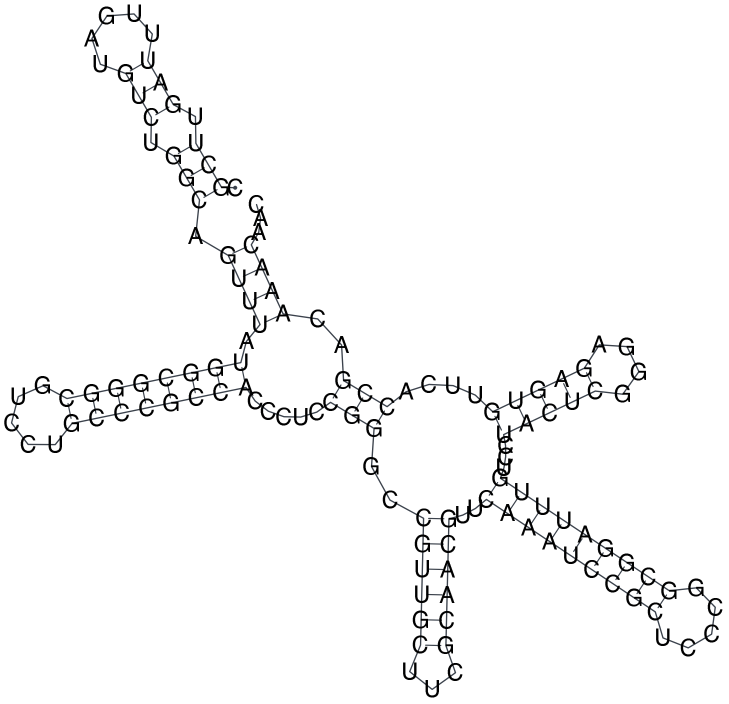

Figure 10: The structure of the fifth ranked prediction (based on the sequence on the forward strand).