Biochemical and Biophysical Research Communications 372 (2008) 346–350

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Construction of two mathematical models for prediction of bacterial sRNA targets

Yalin Zhao, Hua Li, Yanyan Hou, Lei Cha, Yuan Cao, Ligui Wang, Xiaomin Ying *, Wuju Li *

Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Taiping Road 27#, Haidian District, Beijing 100850, China

ARTICLE INFO

Article history:
Received 13 April 2008
Available online 21 May 2008

Keywords:
Prediction of sRNA targets
Support vector machines
Naïve Bayes method
Feature forward selection
Model

ABSTRACT

Accurate prediction of sRNA targets plays a key role in determining sRNA functions. Here we introduced two mathematical models, sRNATargetNB and sRNATargetSVM, for prediction of sRNA targets using Naïve Bayes method and support vector machines (SVM), respectively. The training dataset was composed of 46 positive samples (real sRNA–targets interaction) and 86 negative samples (no interaction between sRNA and targets). The leave-one-out cross-validation (LOOCV) classification accuracy was 91.67% for sRNATargetNB, and 100.00% for sRNATargetSVM. To evaluate the performance of the models, an independent test dataset was used, which contained 22 positive samples and 1700 randomly generated negative samples. The results showed that the classification accuracy, sensitivity, and specificity were 93.03%, 40.90%, and 93.71% for sRNATargetNB and 80.55%, 72.73%, and 80.65% for sRNATargetSVM, respectively. Therefore, the presented models provide support for experimental identification of sRNA targets. The related software and supplementary materials can be downloaded from webpage http://www.biosun.org.cn/srnatarget/.

In bacteria, there exist some small non-coding RNAs (sRNAs) with 40–500 nucleotides in length. For example, more than 70 sRNAs have been found in *Escherichia coli* (*E. coli*) [1,2]. Most of them function as posttranscriptional regulation of gene expression through binding to the translation initiation region (TIR) of their target mRNAs, in which Hfq-protein acts as RNA chaperone [3–10]. In positive regulation [10], the binding region is located in 90–120 nt upstream from the initial start codon, and the role is to activate expression of target genes. In negative regulation, the binding region is near the SD sequence, and the role is to block ribosome binding or trigger degradation of both sRNA and its targets [11]. Therefore, accurate prediction of sRNA targets plays a key role in determining sRNA functions. As the known number of targets in positive regulation is limited, we only took negative regulation into account.

In their recent review [12], Vogel and Wagner systematically summarized the experimental and bioinformatics approaches for identification of sRNA targets. Although sRNA targets should be finally experimentally verified, computational methods still provide a labor-saving way to identify sRNA targets. Up to now, three prediction models have been presented [10,11,13].

In Zhang's model [11], the Smith–Waterman local sequence alignment algorithm was modified by incorporating following information, which included sRNA secondary structure, characteristics of Hfq-binding site, the TIR −35 to 15 of candidate target mRNAs, and conservation profiles of both sRNA and its candidate targets TIR in eight close relatives of *E. coli* K-12. For each sRNA, the score between the sRNA and each mRNA was calculated, and the mRNAs with top scores were considered as the candidate targets. Among 10 verified sRNA–mRNA interactions, there were seven interactions among the top-50 predicted targets. The prediction accuracy was 70.00%. Because the conservation profiles were considered, their algorithm cannot be used to predict targets of species-specific sRNAs or other bacterial sRNAs. In addition, they only considered the secondary structure of sRNA rather than the secondary structure of sRNA–mRNA complex, which maybe generate prediction bias.

In TargetRNA model [10], Tjaden and their colleagues developed two methods for sRNA target prediction, namely, individual basepair method and stacked basepair method. To predict sRNA targets, the hybridization score between sRNA and its candidate targets was firstly calculated using either one of above two methods. Then, the *P* value was obtained by assuming the score abiding by extreme-value distribution. The candidate targets usually had smaller *P* values. Two main related parameters, the size of TIR and the length of seed match, were optimized using 12 validated sRNA–mRNA interactions. The optimal parameters were −30 to

Abbreviations: SVM, support vector machines; *E. coli*, *Escherichia coli*; sRNA, small non-coding RNAs; TIR, translation initiation region; LOOCV, leave-one-out cross-validation.

* Corresponding authors. Fax: +86 10 68213039.
E-mail addresses: yingxm@nic.bmi.ac.cn (X. Ying), liwj@nic.bmi.ac.cn (W. Li).

0006-291X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2008.05.046

20 for TIR and 9 nt for seed match. Among the 12 targets, there were 8 targets correctly predicted. The prediction accuracy was about 66.67%.

In the target prediction methods presented by Cossart group [13], four validated sRNA–target interactions were used to optimize the related thermodynamic parameters, which contained stacking effects, and the cost of bulge and interior loops. These optimized parameters were further applied in calculating the strengths of sRNA–mRNA duplexes. During target prediction, two regions on the mRNA were considered, which were 5′ regions −140–90 (TIR) and 3′ regions spanning 60-nt upstream of the translation stop codon and 90-nt downstream of the stop codon. Finally, the presented method was used to predict the targets of nine novel sRNAs discovered by them, and some predictions were experimentally verified.

In summary, total number of positive samples for the above models is limited. For example, among the three presented models, TargetRNA model had the largest number of samples, only twelve [10]. In addition, different regions around TIR −35–15, −30–20, and −140–90 were used in Zhang's, Tjaden's and Cossart's models, respectively. Then, which region or regions combination was the optimal? To solve these two problems, firstly, we collected 46 positive samples and 86 negative samples as the training dataset. Then, the technique of RNA secondary structure profile was used to find the optimal combinations of different regions [14]. Finally, two models, sRNATargetNB and sRNATargetSVM, for prediction of sRNA targets were constructed using Naïve Bayes method and SVM, respectively.

Materials and methods

Training dataset. To construct models for prediction of sRNA targets, we firstly collected 46 positive samples and 86 negative samples as the training dataset (Supplementary Table 1). Then, we chose the region −80 to 50 (TIR) as the candidate region so that most targets in negative regulation can be considered. Finally, the technique of RNA secondary structure profile was used to search the optimal combinations of regions for prediction of sRNA targets [14].

Test dataset. To evaluate performance of the models, we constructed an independent positive set TESTP and 10 randomly generated negative sets $TESTN_{1-10}$. The TESTP contained 22 positive samples. The $TESTN_{1-10}$ was constructed as follows. For each sRNA involved in positive samples in the training dataset, 10 mRNAs were randomly selected from 4131 mRNAs in *E. coli* (NCBI NC_000913), and 10 sRNA–mRNA interactions were formed. The processes were repeated 10 times, and the sets $TESTN_{1-10}$ were constructed. During random selection of mRNAs, the 132 mRNAs in the training dataset were excluded. Because there were 17 sRNAs involved in the positive samples in the training dataset, the total number of negative samples in the $TESTN_{1-10}$ were 1700 (Supplementary Table2).

Features for model construction. As pointed out in models [10], thermodynamic stability of sRNA–target complex is an important index to describe sRNA–target interaction. But in their models, they only considered one region around TIR at each time. Here we considered multiple regions spanning TIR at the same time. Firstly, we extracted the region −80 to 50 spanning TIR for each potential target. Secondly, we extracted all possible 1000 subsequences around the core region −30 to 30 (Fig. 1). Thirdly, for each subsequence, we concatenated it to sRNA sequence with "LLLLL" between them, and two linking modes, sRNA-LLLLL-subsequence or subsequence-LLLLL-sRNA, were considered [15]. Finally, RNAfold program [16] was used to predict the minimum free energy secondary structure for the above two linked sequences with default parameters, and the structure with lower free energy was used to calculate the following features.

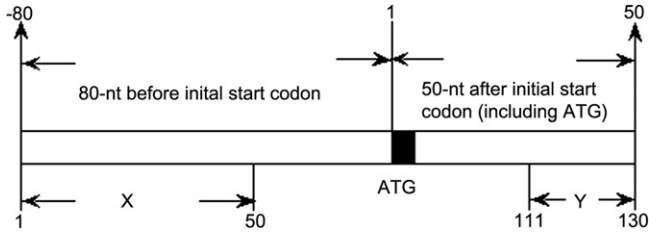

Fig. 1. Demonstration to extract all possible subsequences, correspondent to the interval $[X,Y]$, was given, where $1 \leqslant X \leqslant 50$ and $111 \leqslant Y \leqslant 130$. The total number of subsequences was 1000.

Based on the minimum free energy secondary structure of merged sequence, the percent composition of bases in interior loops, buldge loops, hairpin loops, helical regions, and multibranch loops were calculated, which represented the first to fifth features, respectively (Features 1–5). The sixth feature was the average free energy $\Delta G_m / L_m$, where ΔG_m was the minimum free energy, and L_m was the length of the merged sequence (Feature 6). The seventh feature was the difference of free energy $\Delta G_m - \Delta G_s - \Delta G_T$, where ΔG_m, ΔG_s, and ΔG_T represented the secondary structure free energy of merged sequence, sRNA, and target TIR, respectively (Feature 7). In addition, as demonstrated in TargetRNA model, the length of seed match was an important element in predicting sRNA–mRNA interaction. We took it as the eighth feature (Feature 8). Finally, considering that Hfq-protein usually binds to the A/U rich region, we calculated the percent composition of A + U in single chain, derived from the minimum free energy secondary structure of target TIRs and sRNAs, as the ninth and 10th features (Features 9–10).

Because each target TIR gave 1000 subsequences and each sRNA-subsequence complex provided 10 features, each sRNA–target complex was described using 10,000 features. For 132 sRNA–mRNA interactions in the training dataset, we got a matrix with the size $10{,}000 \times 132$ (Supplementary Table 3). As we did in previous work [14], we called this matrix as the secondary structure profile of sRNA–mRNA interaction. Obviously, the matrix is very similar to the gene expression profile from array technology. Therefore, many methods and tools for gene expression profile could be used. In view that the object of present study was to predict sRNA–mRNA interaction, supervised learning methods were used.

Discriminant analysis. To construct models for prediction of sRNA targets, both Tclass classification system and SVM were used. The Tclass was originally written for gene expression profile-based sample classification [17–19]. In Tclass system, both Fisher and Naïve Bayes methods were integrated with the feature forward selection method. For the present task, the Naïve Bayes method was used.

In addition to Tclass system, we also used SVM [20] to build prediction model using the same training dataset. In recent years, SVM have been successfully applied to some biological problems [15,21].

Results

Standard t-test analysis

Before model construction, the standard *t*-test was used to detect the difference of each feature. The detailed *t*-test results were provided in Supplementary Table 4. There were 4293 features with $P \leqslant 0.01$. The top 40 features ($P \leqslant$ 1E-10) were all the seventh features derived from different subsequences, which showed that the difference of free energy played a key role in prediction of sRNA targets.

Construction of the model sRNATargetNB using Tclass

To construct the model, sRNATargetNB, for prediction of sRNA targets, both Naïve Bayes method and feature forward selection were used with LOOCV classification accuracy as the object function. Tclass system automatically found the optimal combination of features with the number of features from 1 to 20. For each feature number, 10 optimal feature sets were recorded. The relationship between the number of features and LOOCV classification accuracy was displayed in Fig. 2. The highest accuracy was 91.67%. Moreover, the results indicated that many feature sets provided the same classification accuracy. Then, which feature set should be used to construct the model? To solve this problem, the stability analysis was performed. For each optimal feature set, the 132 samples in the training dataset were randomly divided into two groups with partition ratio 75%. The major part was used to construct the classifier, and the minor part was used to calculate classification accuracy. The above processes were repeated 1000 times and the average classification accuracy was taken as the stability index. The relationship between the stability index and the number of features was also given in Fig. 2. From Fig. 2, it can be seen that the highest stable index 0.89 was obtained using six features, which were 198, 567, 1259, 1839, 4102, and 5307, respectively (Table 1). Finally, the feature set was selected as the final best feature set, and the correspondent 1000 classifiers were taken as the final prediction model.

To predict whether a sample was positive or negative, all 1000 classifiers were used. The sample would be predicted to be positive if there were more than 500 classifiers to predict it to be positive. Otherwise, it would be predicted to be negative. Based on the presented model, the 132 samples in training dataset were predicted. There were 35 positive and 86 negative samples correctly predicted. Therefore, the prediction accuracy was 91.67% (121/132), which was higher than 70.00% from Zhang's model, and also higher than 66.7% from program TargetRNA on their training datasets [10]. The sensitivity and specificity were 76.1% (35/46) and 100.00% (86/86), respectively.

Construction of the model sRNATargetSVM using SVM

To construct the model sRNATargetSVM, the LibSVM package (version 2.83) was used [22]. For performance comparison to the model sRNATargetNB, we took LOOCV classification accuracy as the object function. Firstly, we took three feature sets, namely SET1, SET2, and SET3, to construct the models. The SET1 contained all 10,000 features. The SET2 contained 3090 features with $P \leqslant 0.001$. The SET3 included 1785 features with $P \leqslant 0.00001$. Then, for each feature set, the penalty parameter C and the radial basis function (RBF) kernel parameter γ were tuned using the grid search strategy. The related models were marked sRNATargetSVM1, sRNATargetSVM2, and sRNATargetSVM3, respectively. Finally, the prediction results were provided in Table 2, which showed that sRNATargetSVM1 was the best. Compared to the results from sRNATargetNB (91.67%), sRNATargetSVM1 gave the better results.

Performance evaluation of the models

To evaluate performance of the presented models objectively, all four models, sRNATargetNB, sRNATargetSVM1, sRNATargetSVM2, and sRNATargetSVM3, were used to predict the samples in the test dataset. The prediction results were given in Table 3.

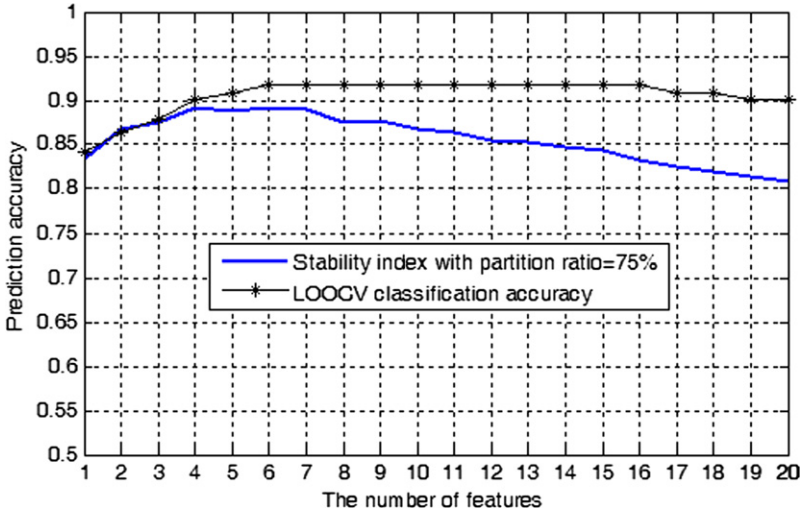

Fig. 2. The relationship between the number of features and LOOCV classification accuracy or stability index was displayed. The stability index was the average of prediction accuracies from the minor parts in 1000 simulations with partition ratio 75%.

Table 1
The meaning of the selected features for sRNATargetNB model

F_t	Interval	TIR	Meaning
198	[1,130]	[−80,50]	The length of seed match between sRNA and mRNA TIR
567	[3,127]	[−78,47]	The difference of free energy $\Delta G_m - \Delta G_s - \Delta G_T$
1259	[7,116]	[−74,36]	The percent composition of A + U in single chain of mRNA TIR
1839	[10,114]	[−71,34]	The percent composition of A + U in single chain of mRNA TIR
4102	[21,121]	[−60,41]	The percent composition of bases in buldge loops
5307	[27,121]	[−54,41]	The difference of free energy $\Delta G_m - \Delta G_s - \Delta G_T$

F_t column stands for the features selected for model construction. See Fig. 1 for detailed information about Interval column and TIR column.

Table 2
The classification accuracy, sensitivity, and specificity from three feature sets using SVM

FeatureSet	N_f	C	γ	Acc (%)	Se (%)	Sp (%)
SET1	10,000	32.0	1.2207×10^{-4}	100.00	100.00	100.00
SET2	3090	2.0	9.7656×10^{-4}	91.67	95.35	84.78
SET3	1785	8.0	9.7656×10^{-4}	84.09	98.84	56.52

N_f column represents the number of features in each feature set. Acc, Se, and Sp stand for classification accuracy, sensitivity, and specificity on the training dataset, respectively.

From the classification accuracy or specificity, sRNATargetNB (Threshold = 1000) gave the best results. The accuracy, sensitivity and specificity were 93.03%, 40.90% and 93.71%, respectively. From the sensitivity, sRNATargetSVM1 provided the best results. The accuracy, sensitivity, and specificity were 80.55%, 72.73%, and 80.65%, respectively. Although sRNATargetSVM1 gave the best sensitivity, we still recommended the model sRNATargetNB (Threshold = 1000) for prediction of sRNA targets because of the following two reasons. First, for each potential sRNA–mRNA interaction, sRNATargetNB model only needed six features. However, sRNATargetSVM1 asked for 10,000 features, and therefore had the longer running time. Second, sRNATargetNB (Threshold = 1000) had high specificity, which generated less number of false positive results and gave less number of targets, which can be more easily verified by experiments.

Comparison to the TargetRNA program

To compare the performance of our models with TargetRNA program, the four sRNA sequences involved in 22 positive samples in the test dataset were input into the TargetRNA web server [10]. Their targets were predicted using the default parameters. The results indicated that there were four interactions detected, which were mica-ompA, RybB-ompN, GcvB-dppA, and GcvB-oppA, respectively. However, there were more than nine interactions detected by our models (except sRNATargetSVM3). Our models provided better performance.

Targets prediction for all known sRNAs in E. coli

For each known sRNA in *E. coli*, the model sRNATargetNB was used to predict its targets. The detailed information was provided in Table 4. From Table 4, we found that the number of targets varied from 8 to 1856 for the Threshold = 500, 4 to 1647 for the Threshold = 800, and 3 to 1055 for the Threshold = 1000, respectively. Obviously, sRNATargetNB (Threshold = 1000) gave the less number of targets. Among all 55 sRNAs, there were 20 sRNAs, 27 sRNAs, and 44 sRNAs with the number of targets less than 51, 100, and 200, respectively. Therefore, the presented models provided support for experimental identification of sRNA targets.

Table 4
The number of predicted targets for all known sRNAs in *E. coli*

sRNA	Target number			sRNA	Target number		
	*Th = 500	Th = 800	Th = 1000		*Th = 500	Th = 800	Th = 1000
RyjC	8	4	3	RyeB	405	296	102
RdlA	77	41	7	MicF	331	256	106
RdlB	68	43	8	CsrC	335	270	127
RdlD	77	42	9	IsrC	427	317	131
SokC	90	49	11	RyfA	470	351	134
RdlC	72	43	11	RyjB	494	377	134
RygC	71	42	11	OmrB	501	384	139
SelC	60	41	13	RyfB	463	345	149
DsrA	110	73	14	MicC	469	355	158
RyfC	129	87	19	RybB	542	403	159
SroB	122	91	20	OmrA	589	446	160
SokB	118	83	21	GadY	570	418	161
RyeC	171	100	21	IsrA	483	379	165
RyeD	109	67	22	PsrO	557	439	178
MicA	107	77	24	RyeE	591	466	178
RygD	135	88	26	IsrB	519	400	181
DicF	191	132	32	RyeA	756	568	210
OxyS	202	138	39	SgrS	740	577	234
RygE	153	108	42	RprA	823	669	274
Spot 42	255	167	50	PsrD	879	694	314
RyjA	209	161	52	RyhB	878	709	316
Tff	323	211	57	RybA	908	716	329
SokA	394	289	73	RyiA	981	803	414
IstR	365	257	87	RyhA	1064	865	437
RydB	354	277	94	ResX	1442	1243	708
SroF	363	256	95	CsrB	1506	1324	822
RydC	406	392	98	GcvB	1856	1647	1055
PsrN	436	311	101				

* Th stands for threshold for target prediction using sRNATargetNB model.

Discussion

Here we presented two models, sRNATargetNB and sRNATargetSVM, for prediction of sRNA targets using Naïve Bayes method and SVM, respectively. The first model consisted of 1000 classifiers derived from six features. The second model was constructed using all 10,000 features. The LOOCV classification accuracy was 91.67% and 100.00% for sRNATargetNB and sRNATargetSVM, respectively. Compared to other models, which gave the classification accuracy 67.7% and 70.0% on the training dataset [10,11], our models gave better results.

To construct models for prediction of sRNA targets, the concept of RNA secondary structure profile was used. It has been successfully applied in the construction of mathematical model for high-level expression of foreign genes in pPIC9 vector [14]. Because we do not know the exact TIRs to interact with sRNA, all possible

Table 3
The performance of the models on the test dataset

Model	TESTP&TESTN$_{1-10}$						
	TP	TN	FP	FN	Acc	Se	Sp
sRNATargetNB (Threshold = 500)	12	1453	247	10	0.8508	0.5455	0.8547
sRNATargetNB (Threshold = 800)	11	1505	195	11	0.8804	0.5000	0.8853
sRNATargetNB (Threshold = 1000)	9	1593	107	13	0.9303	0.4090	0.9371
sRNATargetSVM1	16	1371	329	6	0.8055	0.7273	0.8065
sRNATargetSVM2	13	1103	597	9	0.6481	0.5909	0.6488
sRNATargetSVM3	3	1392	308	19	0.8101	0.1364	0.8188

The same positive test dataset TESTP was combined with each negative test dataset TESTN$_i$ (i = 1,2,...,10). For each sample in each combination TESTP& TESTN$_i$(i = 1,2,...,10), the presented models were used to predict whether it was positive or negative. The sample was predicted to be positive if there were more than 500 classifiers (''Ttreshold = 500'') or 800 classifiers (''Threshold = 800'') or 1000 classifiers (''Threshold = 1000'') to predict it to be positive. Here we want to point out that prediction results for the negative test dataset TESTN$_i$ (i = 1,2,...,10) were merged together. Acc, Se, and Sp stand for classification accuracy, sensitivity, and specificity, respectively.

TIRs around the core region −30–30 were considered. Then, through using the feature forward selection method, we found the optimal feature set with six features (Table 1). From this result, we guess that sRNA–mRNA interaction is a dynamic process, which needs to consider six TIRs to take part in the sRNA–mRNA interaction at the same time. In addition, we also noticed that the feature, the length of seed match between sRNA and mRNA, had been selected during automatic process for feature selection. Through the *t*-test, we got the *t* and *P* values were 3.0830 and 0.0025, respectively. On average, the length of seed match in positive set is longer than that in negative set. This result supported the idea in TargetRNA model, i.e., the length of seed match plays a key role in determining sRNA–mRNA interaction.

During model construction, dataset collection is one of three important parts. The other two parts are machine learning methods and feature selections. Here we want to emphasize two points. First, even though all samples in the training dataset are from *E. coli* K-12, the models can be applied for other bacterial genomes. For example, among 16 sRNA–mRNA interactions in the test dataset from *Salmonella typhimurium* LT2, there were nine samples correctly predicted. Second, since the regulation process of gene expression is very complex, it is very difficult to distinguish primary and secondary targets [12]. For example, some sRNA targets maybe act as transcription factors. If this kind of targets were downregulated, the expression of those genes regulated by these targets will be also downregulated. Obviously, the downregulated genes by the targets are not the real targets of sRNAs. In our training dataset, we assumed that all positive samples are primary targets. With more and more primary targets found, the models will be improved in future.

Acknowledgments

This work was supported by the National High Technology Development Program of China under Grant No. 2006AA02Z323, and National Sciences Foundation of China under Grants Nos. 90608004 and 30470411.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2008.05.046.

References

[1] M.L. Ariane, H.A. Del Portillo, A.M. Durham, Computational methods in noncoding RNA research, J. Math. Biol. 56 (2008) 15–49.
[2] B. Tjaden, Prediction of small, noncoding RNAs in bacteria using heterogeneous data, J. Math. Biol. 56 (2008) 183–200.
[3] N. Delihas, S. Forst, MicF: an antisense RNA gene involved in response of *Escherichia coli* to global stress factors, J. Mol. Biol. 313 (2001) 1–12.
[4] E. Masse, S. Gottesman, A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*, Proc. Natl. Acad. Sci. USA 99 (2002) 4620–4625.
[5] E. Masse, C.K. Vanderpool, S. Gottesman, Effect of RyhB small RNA on global iron use in *Escherichia coli*, J. Bacteriol. 187 (2005) 6962–6971.
[6] S. Altuvia, A. Zhang, L. Argaman, A. Tiwari, G. Storz, The *Escherichia coli* OxyS regulatory RNA represses fhlA translation by blocking ribosome binding, EMBO J. 17 (1998) 6069–6075.
[7] J. Vogel, L. Argaman, E.G. Wagner, S. Altuvia, The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide, Curr. Biol. 14 (2004) 2271–2276.
[8] S. Chen, A. Zhang, L.B. Blyn, G. Storz, MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli*, J. Bacteriol. 186 (2004) 6689–6697.
[9] J. Vogel, C.M. Sharma, How to find small non-coding RNAs in bacteria, Biol. Chem. 386 (2005) 1219–1238.
[10] B. Tjaden, S.S. Goodwin, J.A. Opdyke, M. Guillier, D.X. Fu, S. Gottesman, G. Storz, Target prediction for small, noncoding RNAs in bacteria, Nucleic Acids Res. 34 (2006) 2791–2802.
[11] Y. Zhang, S. Sun, T. Wu, J. Wang, C. Liu, L. Chen, X. Zhu, Y. Zhao, Z. Zhang, B. Shi, H. Lu, R. Chen, Identifying Hfq-binding small RNA targets in *Escherichia coli*, Biochem. Biophys. Res. Commun. 343 (2006) 950–955.
[12] J. Vogel, E.G. Wagner, Target identification of small noncoding RNAs in bacteria, Curr. Opin. Microbiol. 10 (2007) 262–270.
[13] P. Mandin, F. Repoila, M. Vergassola, T. Geissmann, P. Cossart, Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets, Nucleic Acids Res. 35 (2007) 962–974.
[14] B. Wu, L. Cha, Z. Du, X. Ying, H. Li, L. Xu, X. Zheng, E. Li, W. Li, Construction of mathematical model for high-level expression of foreign genes in pPIC9 vector and its verification, Biochem. Biophys. Res. Commun. 354 (2007) 498–504.
[15] S.K. Kim, J.W. Nam, J.K. Rhee, W.J. Lee, B.T. Zhang, miTarget: microRNA target gene prediction using a support vector machine, BMC Bioinformatics 7 (2006) 411.
[16] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (2003) 3429–3431.
[17] W. Li, M. Xiong, Tclass: tumor classification system based on gene expression profile, Bioinformatics 18 (2002) 325–326.
[18] T. Xiao, W. Ying, L. Li, Z. Hu, Y. Ma, L. Jiao, J. Ma, Y. Cai, D. Lin, S. Guo, N. Han, X. Di, M. Li, D. Zhang, K. Su, J. Yuan, H. Zheng, M. Gao, J. He, S. Shi, W. Li, N. Xu, H. Zhang, Y. Liu, K. Zhang, Y. Gao, X. Qian, S. Cheng, An approach to studying lung cancer-related proteins in human blood, Mol. Cell Proteomics 4 (2005) 1480–1486.
[19] W. Li, How many genes are needed for early detection of breast cancer, based on gene expression patterns in peripheral blood cells?, Breast Cancer Res 7 (2005) E5.
[20] C. Cortes, V. Vapnik, Mach. Learn. 20 (1995) 273.
[21] D. Agranoff, D. Fernandez-Reyes, M.C. Papadopoulos, S.A. Rojas, M. Herbster, A. Loosemore, E. Tarelli, J. Sheldon, A. Schwenk, R. Pollok, C.F. Rayner, S. Krishna, Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum, Lancet 368 (2006) 1012–1021.
[22] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machine, 2001. Available from: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.