A Kernel Method for MicroRNA Target Prediction Using Sensible Data and Position-Based Features

Sung-Kyu Kim, Jin-Wu Nam, Wha-Jin Lee, and Byoung-Tak Zhang
Graduate Program in Bioinformatics
Center for Bioinformation Technology (CBIT)
Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University, Seoul 151-744, Korea
{skkim, jwnam, wjlee, btzhang}@bi.snu.ac.kr

***Abstract*—MicroRNAs (miRNAs) are small endogenous RNAs of ~22nt that act as direct post-transcriptional regulators in animals and plants. MicroRNAs generally perform a function by binding to the complementary site on the 3' untranslated region of its target gene and especially the 8mers on the 5' part of miRNA seems important as a seed. Computational methods for miRNA target prediction have been focusing on this seed region, but recent researches revealed that the specificity of the seed region may be sharply decreased even by a point mutation. In this paper, we present a kernel method for miRNA target prediction in animals, which improves the prediction performance with biologically sensible data and position-based features reflecting the way of miRNA:mRNA pairing mechanism. In building a training dataset, we choose experimentally verified data only to improve the quality of dataset by excluding randomly synthesized one and consequently to make the result of learning valid. We use sensitivity, specificity, and area under ROC curve as performance measures of our algorithm and compare the results of various dataset configurations. The overall results were 92.1% in sensitivity, 83.3% in specificity, and 0.931 in area under ROC curve. With position-based features, an increase of 3.3% in sensitivity and 1.6% in specificity were observed. In the feature selection experiment to investigate the role of individual position-based features, the result suggests that pairing at positions 4, 5, 6 of the seed part is of greater importance than the other positions together with the general stability of the seed part.**

I. INTRODUCTION

MicroRNAs are endogenous ~22 nucleotide (nt) non-coding RNAs that act as post-transcriptional regulators in animals and plants. They act by binding to the complementary sites on the 3' untranslated region (UTR) of the target gene to induce cleavage with near perfect complementary to repress productive translation [1-6]. The choice between the translational inhibition and destruction is thought to be governed by the degree of mismatch between miRNA and its target mRNA. Actually, the behavior of miRNA has difference between animals and plants. The miRNAs of plants tend to show near perfect complementarity to their target messenger RNAs (mRNAs) but the miRNAs of animals commonly have imperfect characteristics including mismatches, gaps, G:U wobble pairs and so on [7-13]. That makes it hard to find target mRNAs only with sequence complementarity in animals. Nevertheless, strong sequence conservations observed in target mRNA sites as well as in miRNA sequences make it possible to develop the program for prediction of potential targets [14-16]. This evolutionarily meaningful evidence shows the importance of sequence preservation as a requirement for function. Particularly, no specific role has been explained to the 3' end of miRNAs even though miRNAs tend to be conserved over their whole length.

To date, several target prediction algorithms have been proposed. In plant, similarity based approaches have been shown high performance because of near perfect complementarity between miRNA and its target mRNA [17, 18]. However, the simple similarity based method is not appropriate to animals due to the imperfectness of the miRNA:mRNA match. Recent works in animals have been often based on both the complementarity to the 5' part of miRNAs and conserved motifs over species [10, 19-21], which can be implemented by a model containing weighted position features and comparative information to sense target mRNA sites as well as to reduce false-positives. Scoring methods using dynamic programming [21-23] and complementarity based strategy [19, 24] were preferred to rank the results. They have been quite successful at a few top-ranked results. In our previous work using SVM, we also identified top-ranked potential and known targets of *Caenorhabditis elegans* with biologically meaning [21, 25]. However, they seem to miss considerable amounts of real targets as recent research results found that miRNAs regulate more than 10% of protein production [21, 25]. For example, in case that a real target interaction is human-specific, it is unable to identify the target from other species and lots of false positives will be generated even though they can find putative targets.

Generally, it is natural that the efficiency and the reliability of a machine learning algorithm depend on how to choose specific features and sensible data. However, most of previous works introduced predicted miRNA:mRNA interaction data or randomly generated negative data in order to predict the target genes of miRNA due to both the lack of experimental data and

0-7803-9387-2/05/$20.00 ©2005 IEEE

the difficulties in gathering appropriate data for computational approaches. Though a couple of researches successfully have applied machine learning methods so far [26, 27], they might produce invalid or biased results because of the data problems described above. To overcome these, it is essential to collect biologically meaningful data with reasonable number.

In this paper, we propose a kernel method for miRNA target prediction in animals which learns from the real data. The data depends on several wetlab experiment results to avoid the shortcomings of non-experimental dataset and of complementarity based algorithms. These experiments were for the experimental verification of miRNA target genes and most of the target sites showed high complementarity to the miRNA seed part. Also, we used various feature set including position specific information as input features of the kernel to improve the specificity of algorithm.

A kernel method is popular in modern statistical branch, particularly in probability density estimation and regression function approximation. Aizerman et al. employed radial basis function (RBF) kernels to reduce a convergence proof for the potential function classifier to the linear perceptron case [28]. Bernhard et al. constructed the Support Vector Machines (SVMs) [29], a generalization of the so-called optimal hyperplane algorithm. Actually, the kernels commonly can be used as a similarity measure by automatically providing a vectorial representation of the data in the feature space. In this paper, we have implemented RBF kernel with heterogeneous miRNA target features including position-specific features, transcribed into high dimensional space to apply to SVM classifier.

II. Methods

A. Building Sensible Datasets

In the miRNA function studies, many miRNA target sites have been presented as putative ones based on the complementarity without experimental verification of precise target site [30, 31]. Thus, these data may simultaneously include some actual cases and non-actual binding sites. Consequently, all the non-actual ones should be excluded in order to improve the quality of dataset as long as the exact binding site is not clearly verified by wetlab experiments.

Hence, we decided to collect experimentally verified binding sites of mRNA sequence for several miRNAs over animal species. We carefully collected miRNA:mRNA pair sequence data from the literature. We finally gathered 235 examples including positive and negative data: 152 positives and 83 negatives [1, 10, 11, 23, 32-36]. But the number of negative examples was not enough to learn a classifier. We needed more negative data because negative data contributes to the specificity of a classifier much more significantly. As a matter of fact, specificity is usually more important than sensitivity in genome analysis because slight decrease of specificity value can generate lots of falsely predicted results due to the big size of genome sequence. Fortunately, we could generate 163 artificial negative examples by simple inference. So the total count of examples is 398 (152 positive, 246 negative). Overall procedure is described in the following paragraph.

We noticed that some experiments about knocking out target sites on the target mRNA sequence can give large number of negative examples. In [11], *let-7* miRNA could not repress expression after knocking out target sites of *let-7* miRNA on *lin-41*. And in [4], *let-7* miRNA was inactivated by knocking out target sites on LIN-28. That is, the remaining region on *lin-41* 3'UTR will not work with *let-7* miRNA any more. It is the same to LIN-28. We can now clearly conclude that if all the actual binding sites on *lin-41* and LIN-28 are masked, then all the other remaining sites with favorable seed pairing are apposite to negative examples. In practice, we collected examples with more than 4mer matches at their seed part and discarded the rest in order to improve the quality of dataset. As a result, we got total 163 artificially made negative examples, 50 from *lin-41* and 113 from LIN-28.

Additionally we built the negative dataset consisting of 5000 randomly generated examples, used in several studies, for comparison [19, 21, 24]. The sequence frequencies to generate the random mRNA sequence were the same as in [24], (p_A = 0.34, p_C = 0.19, p_G =0.18, p_U = 0.29). The random examples were selected in the same criterion of the artificial ones. Using randomly generated data as a negative training example [26, 27] is very dangerous because the random data might be quite different from actual data and the specificity can be hyper-estimated.

B. Feature Design

Fundamentally, all features are based on the RNA secondary structure prediction result by RNAfold program in Vienna RNA Package [37]. The general scheme of miRNA and its target mRNA interaction is illustrated in Fig. 1a. RNAfold requires a single linear RNA sequence as input, so the 3' end of the target mRNA sequence and the 5' end of miRNA sequence were connected by the linker sequence, 'LLLLLL'. The 'L' does not mean an RNA nucleotide, so it does not match with any nucleotide and does prevent mRNA and miRNA nucleotide from binding with sequence-specific linker sequence as used in

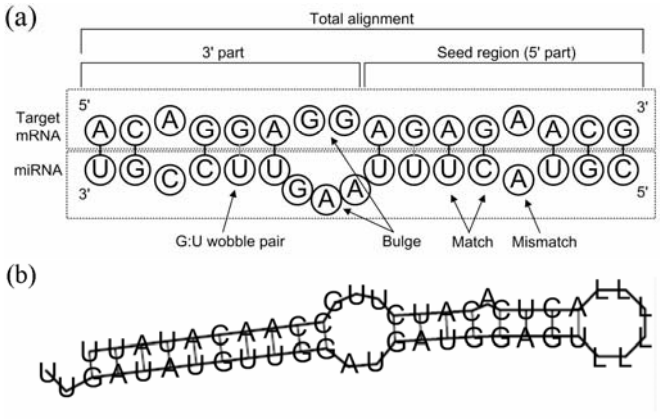

Fig. 1. (a) General scheme of miRNA:mRNA interaction. (b) An example of the secondary structure alignment by RNAfold program, which is a result between the target sequence in the 3'UTR of *lin-41* and *let-7a* miRNA.

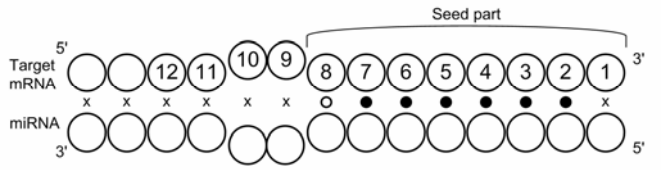

Fig. 2. The importance of position-based pair near the seed part [1]. The black circles indicate very important positions where a point mutation destroys miRNA function. The white circles denote medium important positions where a point mutation either affects or not. The scissor marks stand for non-significant positions regarding point mutation. Indices for position-based features are numbered on the mRNA nucleotide.

[10]. As a result, RNAfold program produces an RNA secondary structure alignment with a linker sequence as shown in Fig. 1b.

The three categories of features were extracted from the dataset based on the secondary structure alignment result by RNAfold program. They consist of position-based, structural, thermo-dynamic features. Position-based features are introduced first in our study, while the other structural and thermo-dynamic features have been widely used in previous studies. We designed 33 features from this procedure.

Position-based features are important features which show the shape and the mechanism of seed pairing. As we stated before, Doench et al. and Brennecke et al. focused on the sequence specificity of miRNA:mRNA pair. They found that a single point mutation could inhibit the miRNA function according to the position of mutation. In contrast to our earlier belief, their research revealed that examples with favorable thermo-dynamic free energy might not regulate expression. So we wondered that what was going to be the real key of the binding mechanism. Position-based features correspond to the way of point mutations on the above two experiments. Fig. 2 shows how the position-based features are organized. We extracted 12 features by numbering from position 1 to 12. The rest positions were ignored because they seemed to be less significant than seed positions [19]. Each position can have 5 states consisting of GC match, AU match, GU match, mismatch and gap. To learn this feature 5 states are translated into decimal values from 1 to 5, respectively.

Next, for structural features, we divided the secondary alignment into 3 parts consisting of 5' part, 3' part and total alignment as shown in Fig. 1a. Each count value of match, mismatch, GC match, AU match, GU match and gap from each 3 part was considered. Total 18 features were extracted as structural features.

Finally, the thermo-dynamic features, free energy value of the 5' part, the 3' part and the total of miRNA:mRNA structure, were obtained from RNAfold program. For the 5' part (seed) of the whole structure, we sliced 8 pairs of the aligned structure beginning from the rightmost position. And the remaining part meant the 3' part. We reassembled each partial sequence with a linker sequence in order to use as an input of RNAfold program. Then RNAfold calculates free energy for each single input strand.

C. Kernel Method as a Learning Algorithm

We used a Support Vector Machine (SVM) to learn the target mRNA discriminating rule from the training dataset. Currently, an SVM is the most prevalent method as a classifier because of its powerful performance and largely applicable features. SVMs, with their string theoretical roots, are known as excellent algorithms for solving classification problems. SVMs allow an implicit mapping of the sample vectors into a high-dimensional, non-linear feature space, in which the samples may be better separated through the use of a similarity function between pairs of samples, called kernel. In order to implement a kernel method, we define several symbols as follow.

Let us denote by $S=\{x_1,\ldots,x_n\}$ a set of miRNA target data to be trained. We suppose that each data x_i is an element of a set X, all possible target data. In order to design data classification method, the dataset S is then represented as the set of features, $\Phi(S)=\{\Phi(x_1),\ldots,\Phi(x_n)\}$, where $\Phi(x)$ can be defined as a real-valued vector. In our study, the size of vector is 33 as mentioned above. The classification method is designed to process a set of pairwise comparisons of dataset x_i and x_j. It is represented by the $n \times n$ matrix of pairwise comparison $k_{i,j}=k(x_i, x_j)$. The $n \times n$ matrix is used as input data of our kernel.

Our SVM algorithm uses RBF kernel with one parameter γ as described in the equation 1.

$$\boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp(-\gamma \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2) \qquad (1)$$

In our study, we search an optimal γ, where produces the best result, from 0.01 to 2.0 increasing by 0.01. And the remaining parameters were kept default. We implemented a modified version of the SVM^{light} [38] to meet our problem and to calculate performance evaluation measures.

D. Evaluation Measures

A 10-fold cross-validation is generally known to create a good estimate of the predictive accuracy of classification methods. Averaging accuracies from the 10 individual experiments estimates the predictive accuracy of a classifier. In case of small data, 10 fold cross-validation performs better than 5 fold cross-validation [39]. In our experiments, we use 10 fold cross-validation for both accuracy and parameter estimation.

The sensitivity and specificity are the most popular performance measures of an algorithm. But if these values vary with parameters, we need to choose appropriate thresholds according to the purpose of the algorithm. As we mentioned earlier, specificity is more significant than sensitivity in our study. We carefully tuned kernel parameter to get the best result and the specificity was a more important criterion.

The receiver operating characteristic (ROC) curve is a statistical tool for describing the performance of a test [40]. For binary classification problem there are two kinds of potential errors: false-positive, where the test shows positive values for data that are in fact negative, and false-negative, where the test shows negative values for data that are in fact positive. ROC analysis provides a numerical and graphical representation of the tradeoff between false-positive rates and true-positive rates

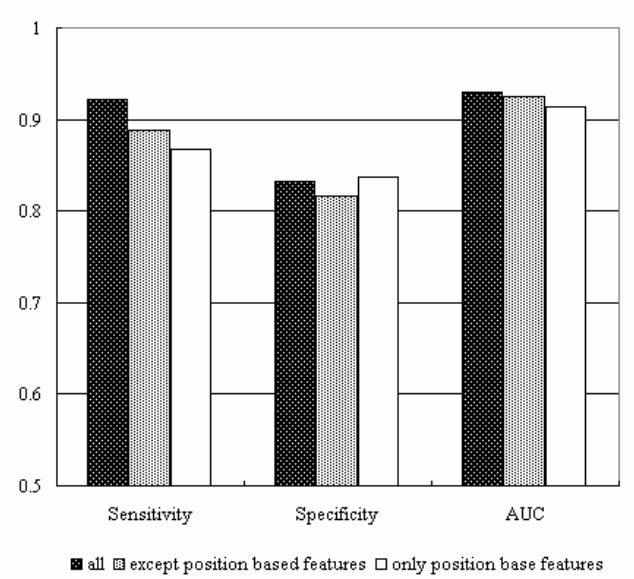

Fig. 3. The performance of the classifier according to the training dataset configuration. The performance is presented in terms of statistical measures: sensitivity = TP/(TP+FN); specificity=TN/(TN+FP) where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

that are produced by the test and allows estimation of optimal cutoff levels for discrimination and filtration of test results. Also an ROC curve can usually be characterized by the area under the curve (AUC). The AUC gives a good indication for the overall performance of classifier and whether one classifier performs better than the other classifier [41]. The bigger the value is, the better the algorithm is. An area of 1.0 is the maximum which means perfect classification and 0.5 means random guess.

III. Results

A. Performance of the Classifier

Fig. 3 shows the performance of our classifier for various feature configurations. First, we tested the classifier with all feature set and the results were 92.1% in sensitivity, 83.3% in specificity and 0.931 in AUC, where F-measure (described as below) is highest. In order to investigate the effect of position-based features, we evaluated the efficiency of the classifier after excluding the 12 position-based features. The efficiency was overall decreased: sensitivity decreased 3.3%, specificity decreased 1.6% and AUC decreased 0.5%. Next, we tested only with position-based features to measure the significance of structural and thermo-dynamic features. Sensitivity decreased more than 5% but specificity remained similar. From this result, we can suppose that the position-based features are main contributors to alleviate sensitivity and specificity, but other features have more effects on the sensitivity.

B. Comparison of Original and Random Dataset

Previous researches often introduced examples randomly

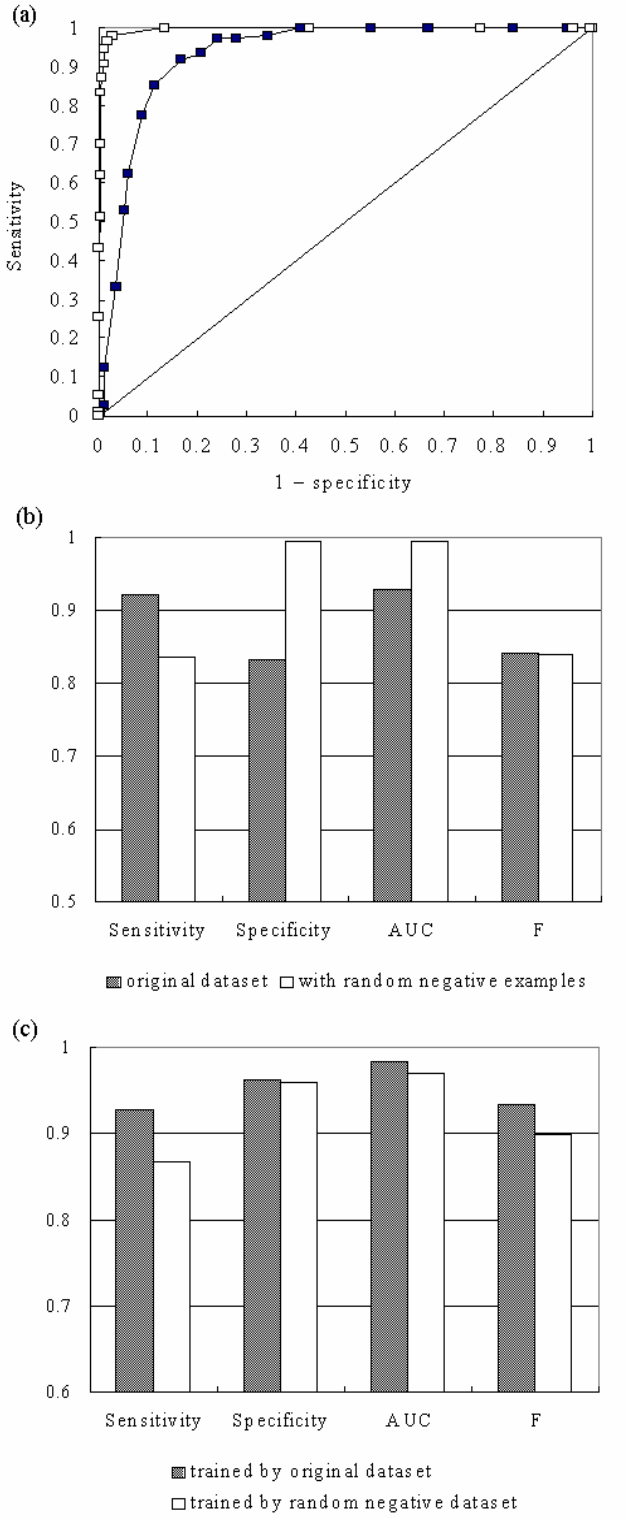

Fig. 4. (a) The ROC curve, which is defined as a plot of test sensitivity as the y-coordinate, versus the false positive rate (FPR; 1–specificity) as the x-coordinate. The black box line is the curve without the randomly generated negative examples whereas the empty box line is the curve with the randomly generated ones. The area under the ROC curve is 0.931 and 0.996, respectively. (b) F-measure comparison between original dataset and dataset with randomly generated negative examples. F-measure=2×(precision×recall)/(precision+recall) where precision=TP/(TP+FP) and recall=sensitivity=TP/(TP+FN). (c) Comparison of classifier efficiency with different negative training dataset.

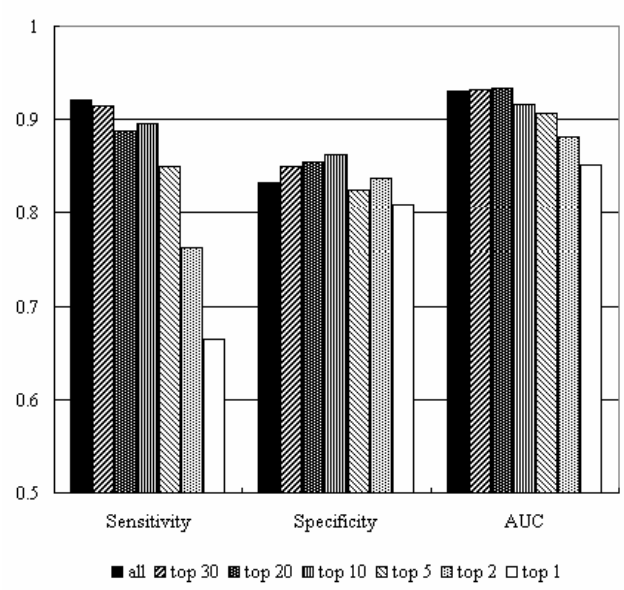

Fig. 5. Performance change according to the number of features

TABLE I
THE TOP 20 CONTRIBUTING FEATURES

Rank	Rank score	Features
1	81.9	position 5
2	79.6	5' part free energy
3	79.1	position 6
4	78.9	position 4
5	78.9	AU match at 5' part
6	77.6	mismatch at 5' part
7	76.6	match at 5' part
8	73.9	total GU match
9	73.4	position 7
10	72.9	position 2
11	71.4	GU match at 5' part
12	70.8	GU match at 3' part
13	70.3	total AU match
14	68.8	position 3
15	68.6	total free energy
16	68.3	position 8
17	67.6	total mismatch
18	66.6	GC match at 5' part
19	65.6	position 9
20	64.3	position 10

generated with background base composition to produce negative training examples [26, 27].

We have performed a comparison of algorithmic efficiency between in using negative data including the random dataset and in using only original negative data. Fig. 4a shows two ROC curves and indicates their performances under given negative dataset, respectively. Contrary to our expectation, the latter seems like more efficient than using original data. This biased result can be explained by two main reasons.

The first, randomly generated negative data often are not biologically feasible and can be easily classified from positive data, and it results in high specificity as in Fig. 4a, b. The second reason is the problem of negative data size. The size of random negative dataset is 5000 though the original negative dataset is 246. The size of random negative dataset becomes almost 20 times. It means that the dataset configuration is seriously biased due to the big size of the random negative data, so the shown result is not reliable one. To make it sense, other reliable measurements which are stringent to biased configuration should be considered.

Consequently, we introduced F-measure [42] focusing on the biased environment. F-measure is the harmonic mean of precision and recall. So it is maximized when precision and recall are similarly high. By drawing F-measure together, we finally concluded that randomly generated data made the classifier significantly biased. White bars in Fig. 4b indicate the performance of our classifier with random negative data. You can find that the F-measure values of the two datasets are almost the same, while the others vary as shown in Fig. 4b. The result with random negative data is almost perfect in specificity and AUC except sensitivity. This makes us wonder how this data can result in such a good specificity and AUC in spite of relatively low sensitivity.

To further investigate this, we next trained our classifier under two different negative dataset configurations as training datasets. The two datasets were original positive dataset with randomly generated negative dataset and with original negative dataset, respectively. That is, they are different only in negative dataset. Test dataset was the original one consisting of the original positive and negative dataset. We did not use 10 fold cross-validation this time since the training datasets were different while using the same test dataset. Finally, the performance of the classifier with random negative data shown in Fig. 4c is not as excellent as the value shown in the previous figure. All measures are lower than the value of original dataset. In [26, 27], they used random negative dataset as negative examples for training. We showed that their results were not optimal and can be improved by including experimentally verified negative examples in the training set.

C. Contribution of Individual Features

We use feature selection to investigate which feature plays more dominant role in miRNA target regulation. Feature selection approaches are used for the following purposes. First, improving the performance of the classifier. Second, producing a cost-effective classifier. And last, understanding the problem better. In our case, the objective is to understand the hidden mechanism of miRNA function. We expected that there must be dominantly functioning features or non-informative features.

We used the Weka [43] for a feature selection tool because it is very easy to use and provides lots of algorithms. Features were evaluated by OneR classifier and Ranker method, and the top 20 contributing features are shown in Table 1.

As might be expected, position-based features are ranked half of the top 10 features. Moreover, position 5 is in the lead. And the consequence pairing of positions 4, 5, 6 may be

important for miRNA function, because they are ranked at 4, 1, and 3, respectively. G:U wobble pair also plays an important but negative role. Usually a G:U pair is known as a disturbing factor [4, 11]. Therefore GU related features were ranked high. More than 3 G:U wobble pair can cause serious degradation of miRNA function [35], which is consistent with this result.

Next time, we wondered that how many features were really contributing to prediction. We prepared datasets consisting of top 1, 2, 5, 10, 20, 30 features, respectively. Each dataset was trained separately and the result is shown in Fig. 5. When we trained the classifier with lowing up to the top 10 features, there was almost small difference from the original in sensitivity. However, using less than top 10 features gives distinctive performance degradation. In contrast to sensitivity, specificity did not vary much over the tested datasets and it remained similar. That is, the specificity has a couple of crucial requirements ranked at the top, while the sensitivity was determined by various factors. This means that the target sites have both common characteristics and special ones which need to be investigated later.

IV. Discussion

In this paper, we proposed a kernel based approach to predicting miRNA target sites with biologically sensible data and investigated which features contribute significantly to the miRNA:mRNA functioning mechanism. An SVM classifier, as a kernel method, performed well on our dataset and produced considerable results. Position-based features contribute to overall performance improvement and, especially, increase sensitivity. Features on positions 4, 5, 6 seem to have more significant effects on seed pairing according to the result of feature selection analysis. This comes with the general belief that the 5' part of paired structure should be stable, and this surely specifies individual positions of importance in the seed.

In this research, we have expected to get higher specificity than 0.9 at least. As we mentioned before, higher specificity is required in genome research because of the big size of genomes. Ours was around 0.83 and it was not that high enough. Thinking of the combination of nucleotide pair, the rate of match and mismatch is 0.25 and 0.75, respectively. Thus, the expected specificity would be 0.75 with one position-based feature where a match is required, position 5 for example. In that sense, our classifier needs to be improved in specificity by introducing more artificial or experimental negative examples in the future.

According to the analysis of [1], there would be three classes of miRNA target sites consisting of canonical 5' dominant, seed dominant, 3' compensatory. Two of these classes need to have strong complementarity on the seed. In 3' compensatory class, however, it needs to have only moderate level of complementarity in the seed, while the 3' part has considerable pairing. Our result failed to explain this. You can find only 1 feature ranked at 12^{th}, which is GU match at 3' part, and almost all of the others are about 5' part. This may be because the result is still biased toward the effect of miRNA seed. Recent studies often concentrate on the conservation of seed motifs [36, 44] and wetlab experiments are performed accordingly. So the experiments about 3' part are rarely done and it makes us hard to get appropriate data to investigate the effect of 3' part.

In addition to this consideration, multi-class classification algorithms may explain this situation. If there really are three distinct classes in miRNA:mRNA pairing mechanism, our binary classification approach will be challenged by strong limitations. Relatively low specificity seems to imply the misleading point of computational prediction algorithms. The lack of dataset is going to be another main limitation. Our dataset is still small to apply machine learning approaches. But in multi-class problems, dataset size should be much larger than binary problem accordingly.

In summary, our approach improved the performance of miRNA target prediction with position-based features. And we showed that the specific position of importance in seed pairing. But there are still limitations on applying computer based approaches. First, the mechanism of miRNA function is not clear. Second, biologically reliable data is scarcely found. Lastly, biological mechanism can be specific regarding species. We believe that our approach produces meaningful result, but it needs to be improved in order to overcome the limitations described above. It will greatly improve the performance of classifier to create more reliable features reflecting the unknown mechanism of miRNA.

Acknowledgment

This work was supported by the National Research Laboratory program (M1041200095-04J0000-03610) of KOSEF and the BK21-IT program of the Korean Ministry of Education. The ICT at Seoul National University provided research facilities for this study.

References

[1] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of microRNA-target recognition," *PLoS Biol.*, vol. 3, pp. e85, 2005.

[2] E. C. Lai, "microRNAs: runts of the genome assert themselves," *Curr. Biol.*, vol. 13, pp. R925-36, 2003.

[3] J. C. Carrington and V. Ambros, "Role of microRNAs in plant and animal development," *Science*, vol. 301, pp. 336-8, 2003.

[4] P. Nelson, M. Kiriakidou, A. Sharma, E. Maniataki, and Z. Mourelatos, "The microRNA world: small is mighty," *Trends Biochem. Sci.*, vol. 28, pp. 534-40, 2003.

[5] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, pp. 281-97, 2004.

[6] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, pp. 350-5, 2004.

[7] C. Llave, Z. Xie, K. D. Kasschau, and J. C. Carrington, "Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA," *Science*, vol. 297, pp. 2053-6, 2002.

[8] J. F. Palatnik, E. Allen, X. Wu, C. Schommer, R. Schwab, J. C. Carrington, and D. Weigel, "Control of leaf morphogenesis by microRNAs," *Nature*, vol. 425, pp. 257-63, 2003.

[9] G. Tang, B. J. Reinhart, D. P. Bartel, and P. D. Zamore, "A biochemical framework for RNA silencing in plants," *Genes & Dev.*, vol. 17, pp. 49-63, 2003.

[10] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen, "Identification of Drosophila MicroRNA targets," *PLoS Biol.*, vol. 1, pp. E60, 2003.

[11] M. C. Vella, K. Reinert, and F. J. Slack, "Architecture of a validated microRNA::target interaction," *Chem. Biol.*, vol. 11, pp. 1619-23, 2004.

[12] M. C. Vella, E. Y. Choi, S. Y. Lin, K. Reinert, and F. J. Slack, "The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR," *Genes & Dev.*, vol. 18, pp. 132-7, 2004.

[13] H. Robins, Y. Li, and R. W. Padgett, "Incorporating structure to predict microRNA targets," *Proc Natl. Acad. Sci. U S A*, vol. 102, pp. 4006-9, 2005.

[14] U. Ohler, S. Yekta, L. P. Lim, D. P. Bartel, and C. B. Burge, "Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification," *RNA*, vol. 10, pp. 1309-22, 2004.

[15] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, pp. 15-20, 2005.

[16] Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl, and H. Margalit, "Clustering and conservation patterns of human microRNAs," *Nucleic Acids Res.*, vol. 33, pp. 2697-706, 2005.

[17] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel, "Prediction of plant microRNA targets," *Cell*, vol. 110, pp. 513-20, 2002.

[18] M. W. Jones-Rhoades and D. P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," *Mol. Cell*, vol. 14, pp. 787-99, 2004.

[19] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, pp. 787-98, 2003.

[20] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in Drosophila," *Genome Biol.*, vol. 5, pp. R1, 2003.

[21] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human MicroRNA targets," *PLoS Biol.*, vol. 2, pp. e363, 2004.

[22] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes," *RNA*, vol. 10, pp. 1507-17, 2004.

[23] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou, "A combined computational-experimental approach predicts human microRNA targets," *Genes & Dev.*, vol. 18, pp. 1165-78, 2004.

[24] N. Rajewsky and N. D. Socci, "Computational identification of microRNA targets," *Dev. Biol.*, vol. 267, pp. 529-35, 2004.

[25] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs," *Nature*, vol. 433, pp. 769-73, 2005.

[26] W. J. Lee, J. W. Nam, S. K. Kim, and B. T. Zhang, "Identification of *Caenorhabditis elegans* MicroRNA Targets Using a Kernel Method " *Genomes & Informatics*, vol. 3, pp. 15-23, 2005.

[27] O. Saetrom, O. Snove, Jr., and P. Saetrom, "Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms," *RNA*, 2005.

[28] E. B. a. L. R. M. Aizerman, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821-837, 1964.

[29] E. B. Bernhard, M. G. Isabelle, and N. V. Vladimir, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. Pittsburgh, Pennsylvania, United States: ACM Press, 1992.

[30] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, pp. 843-54, 1993.

[31] B. Wightman, I. Ha, and G. Ruvkun, "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans," *Cell*, vol. 75, pp. 855-62, 1993.

[32] R. J. Johnston and O. Hobert, "A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans," *Nature*, vol. 426, pp. 845-9, 2003.

[33] S. Yekta, I. H. Shih, and D. P. Bartel, "MicroRNA-directed cleavage of HOXB8 mRNA," *Science*, vol. 304, pp. 594-6, 2004.

[34] P. T. Nelson, A. G. Hatzigeorgiou, and Z. Mourelatos, "miRNP:mRNA association in polyribosomes in a human neuronal cell line," *RNA*, vol. 10, pp. 387-94, 2004.

[35] J. G. Doench and P. A. Sharp, "Specificity of microRNA target selection in translational repression," *Genes & Dev.*, vol. 18, pp. 504-11, 2004.

[36] E. C. Lai, B. Tam, and G. M. Rubin, "Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs," *Genes & Dev.*, vol. 19, pp. 1067-80, 2005.

[37] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucleic Acids Res.*, vol. 31, pp. 3429-31, 2003.

[38] J. Thorsten, "Making large-scale support vector machine learning practical," in *Advances in kernel methods: support vector learning*: MIT Press, 1999, pp. 169-184.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," presented at In the International Joint Conference on Artifcal Intelligence, 1995.

[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29-36, 1982.

[41] A. Shivani, G. Thore, H. Ralf, H.-P. Sariel, and R. Dan, "Generalization Bounds for the Area Under the ROC Curve," *J. Mach. Learn. Res.*, vol. 6, pp. 393-425, 2005.

[42] C. Van Rijsbergen, *Information Retrieval*. London: Butterworths, 1971.

[43] I. H. W. a. E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition ed. San Francisco: Morgan Kaufmann, 2005.

[44] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals," *Nature*, vol. 434, pp. 338-45, 2005.