Recognizing Drosha processing sites by a two-step prediction model with structure and sequence information

Xingchi Hu, Yanhong Zhou*
Hubei Bioinformatics and Molecular Imaging Key
Laboratory, College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, China
E-mail: gaberil.0113@gmail.com, yhzhou@hust.edu.cn

Chuang Ma
School of Plant Sciences
University of Arizona
Tucson, USA
E-mail: chuangma2006@gmail.com

***Abstract*—Drosha is a class of RNase III enzyme plays important roles in the microRNA (miRNA) generation by cleaving primary miRNAs to release hairpin-shaped miRNA precursors. Accurately predicting the Drosha cleavage positions (i.e., processing sites) is helpful for the identification of miRNAs and the understanding of miRNA biogenesis mechanisms. In this study, we presented a Drosha processing site predictor, termed DroshaPSP, with a two-step prediction model by integrating structure and sequence features. Testing results on the Drosophila melanogaster miRNA data showed that DroshaPSP obtained a sensitivity of 0.859, a specificity of 0.999, and a Matthew's Correlation Coefficient of 0.864. We also found that the Shannon entropy is a powerful structure feature for DroshaPSP to distinguish true Drosha processing sites from the nearby pseudo processing sites effectively.**

Keywords-miRNA; Drosha; Shannon entropy; SVM.

I. INTRODUCTION

MicroRNAs (miRNAs) are a large class of ~22nt long non-coding RNAs with important functions that regulate the expression of a large number of target genes in animals and plants [1-2].

Drosha is a member of RNase III family. In animals, except a few miRNAs are processed by miRtron pathway, long primary-miRNAs (pri-miRNAs) are usually cleaved into ~70nt precursor miRNA (pre-miRNA) hairpins by the Drosha enzyme [3-4], and subsequently cut by the Dicer enzyme, which determines mature miRNA and miRNA star. As Dicer processing sites locate closely to Drosha processing sites with the distance about 22bp [5], the Drosha process is thus an important step of determining mature miRNAs. Moreover, the Drosha process also controls the specificity and efficiency of miRNA expression [6]. Therefore, accurately identifying the Drosha processing sites will be extremely helpful for identifying miRNAs, and understanding Drosha process and miRNA biogenesis.

Both experimental and computational methods have been applied to identify Drosha processing sites. By using the tiling microarrays technology, Kadener and colleagues identified 137 Drosha processing region of Drosophila [7]. Due to the advantages of the low-cost and high-speed, computational methods have also been attracting great attentions. The 'Microprocessor SVM' has been developed to predict Drosha processing sites in human genome with the feature set including structure information and base pair information of pre-miRNA [8]. However, the prediction rate for known human 5' Drosha processing sites is 50% approximately. The possible reason is the ignorance of chemical dynamical features of pre-miRNA folding. The Shannon entropy is a recently presented measure characterizing the structure feature in the folding processing of non-coding RNA sequences (ncRNAs) [9]. It would also be important for the Drosha processing step in the miRNA generation [10]. In this study, we developed a new computational method, named DroshaPSP, to predict Drosha processing sites by considering the sequence and structure features including the Shannon entropy. Test results showed that DroshaPSP achieved relatively high prediction accuracy on the Drosophila malanogaster miRNA data. We found that the Shannon entropy plays an important role in Drosha processing site prediction.

II. MATERIALS AND METHODS

A. Data

We chose Drosophila melanogaster as our study species because the Drosha processing sites of miRNAs have been extensively annotated. Here the Drosha processing sites are defined sites as the 5' ends of mature miRNAs or miRNA stars in 5' stem annotated by miRBase [11], if there is no such annotation we assumed that 3' ends of mature miRNAs give a 2nt overhang relative to 5' Drosha processing site, similar with the definition in [8].

The miRNA annotation data of Drosophila melanogaster including pre-miRNA sequences, miRNA hairpin structure, mature miRNA and miRNA star were obtained from miRBase. Of note, the miRNAs processed by miRtron pathway were removed. The genomic sequences of Drosophila melanogaster were downloaded from Ensemble database [12].

B. Architecture of DroshaPSP

DroshaPSP predicts the Drosha processing sites with a two-step prediction model. For each step, the prediction model was constructed with a support vector machine (SVM) classifier [13] with a RBF kernel implemented with the LIBSVM package [14]. For a given DNA sequence, DroshaPSP firstly implements the first SVM classifier

*The corresponding author. Yanhong Zhou, Ph.D., Professor, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Department of System Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China. Tel: +86 27 87792217. Email: yhzhou@hust.edu.cn.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

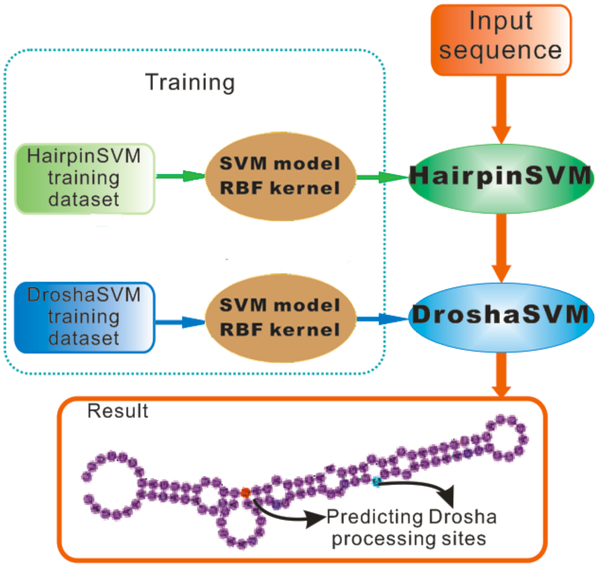
Input sequence
Training
HairpinSVM training dataset
SVM model RBF kernel
HairpinSVM
DroshaSVM training dataset
SVM model RBF kernel
DroshaSVM
Result
Predicting Drosha processing sites

Figure 1. The system architecture of DroshaPSP.

named HairpinSVM to identify the pre-miRNA-like hairpin structures, which are then processed by the other SVM classifier termed DroshaSVM to determine the Drosha processing sites from the hairpin structures (Fig. 1).

1) HairpinSVM: Determining the pre-miRNA-like hairpin structures

There are 12 structure features were integrated into HairpinSVM to determine the most possible hairpin structure of pre-miRNA (Table I). For training this SVM classifier, 641 positive samples were collected from miRNA hairpin structure in miRBase, 3,024 negative samples were hairpin structure given out by folding fragment over 50nt around confirmed pre-miRNA (within 180nt) with RNAfold [Ref]. The optional combination of C and gamma in SVM classifier was selected by a grid search approach with exponentially growing the values of these two parameters.

2) DroshaSVM: Drosha processing site classifier

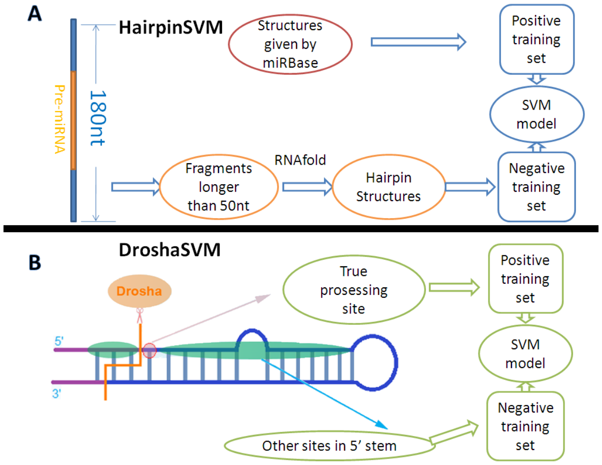

Figure 2. The flow char of constructing prediction models for HairpinSVM (A) and DroshaSVM (B)

TABLE I. THE FEATURES USED IN HAIRPINSVM

ID	Name	Description
1	Length	The length of the sequence
2	Loop_length	The loop size of hairpin structure
3	Stem_length	The stem length of hairpin structure
4	Pair	The number of base pairs in folding result
5	Pair_frac	The fraction of paired base in sequence
7	Insert_count	The number of bulges in the folding structure output by RNAfold
6	Insert_frac	The average length of bulges in sequence
8	Insert_count_frac	The ratio between the nucleotides in bulges and those in the sequence
9	Mfe	The minimal free energy output by RNAfold
10	Ensemble_fe	The free energy of the thermodynamic ensemble
11	Ensemble_fq	The probability of this single structure in the Boltzmann weighted ensemble of all structures.
12	Ensemble_div	The ensemble diversity is the average base-pair distance between all structures in the thermodynamic ensemble.

TABLE II. THE FEATURES USED IN DROSHASVM

ID	Name	Description
1	Loop_Distance	Distance from processing site candidate to loop of the hairpin structure.
2~11	Structure	Structure description of the candidate site and 9nt sites forward are paired or not.
12~21	Base	The base types of the candidate site and 9nt sites forward.
22~31	Probability	The base pairing probability of the candidate site and 9nt sites forward.
32~41	Entropy	The Shannon entropy of the candidate site and 9nt sites forward.

Candidate site is defined as Position_0, other sites are defined as Position_i, where i is the distance from the candidate

DroshaSVM outputs a probability for each candidate Drosha processing site, which is the site at the 5' stem of hairpins output by the HairpinSVM (Fig. 2B). To train DroshaSVM, we collected 641 experimentally validated miRNA or miRNA star's 5' end as positive samples from miRBase database. Other 30,873 sites in 5' stems were considered as negative samples.

C. Performance estimation

The 5-fold cross-validation was used to test the performance of both classifiers. The performance of each prediction model was estimated by five measures: accuracy (ACC), sensitivity (Sn), specificity (Sp), precision (P) and Matthews correlation coefficient (MCC), defined as follow:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})(\mathrm{FN} + \mathrm{FN})}}$$

where TP, FP, TN and FN represent the counts of true positive, false positive, true negative, false negative respectively. The MCC is used to determine the default threshold for both HairpinSVM and DroshaSVM [15].

Besides these measures, we also used the ROC (receiver operating characteristic) curve to comprehensively estimate the performance of two SVM classifiers, which can be quantified with the area under the ROC curve (AUC). The AUC values range from 0.0 to 1.0. The higher the AUC is, the better the classifier performance.

For DroshaPSP program, all the sites in 5' arms of Drosophila melanogaster pre-miRNA hairpins annotated by miRBase formed testing dataset, in which 5' ends of mature miRNAs or miRNA stars produced by Drosha are considered as positive data. The performance is also accessed by the evaluating index of ACC, SN, SP, P and MCC.

III. Result

A. *Performance of the classifiers*

The ROC curve analysis shown the AUC values are 0.964 and 0.974 for HairpinSVM (Fig. 3A) and DroshaSVM (Fig. 3C), respectively. To check the stability of the performance of these two classifiers on the unbalanced training set, we plot the curves of MCC varied with the threshold for HairpinSVM (Fig. 3B) and DroshaSVM (Fig. 3D). These results indicated both classifiers are highly stable. The threshold with the highest MCC is selected as default thresholds, then the performance measurements of HairpinSVM and DroshaSVM were calculated (Table III).

B. *Performance of the DroshaPSP program*

After determining the optimal kernel parameters and threshold values for two SVM classifiers, we further tested the performance of DroshaPSP with testing dataset. The test results showed that ACC, P, MCC and SN can respectively reach 0.998, 0.970, 0.864 and 0.859, when SP was 0.999.

C. *Estimated importance of the features used in classifiers*

To figure out the importance of each feature used in classifiers, the F-score method is applied [16]. The larger the F-score is, the more likely this feature is discriminative. The F-score values of features used in HairpinSVM and DroshaSVM are respectively shown in Fig. 4A and Fig. 4B. We found that the energy feature is one of the most effective features in the selection of pre-miRNA-like hairpins (Fig. 4A). For DroshaSVM, the F-score of the probability, the structure and the Shannon entropy are higher than that of the Base. Moreover, F-sores of the Position_3 to Position_9 are much higher than those of other positions, indicating the importance of these positions in the identification of Drosha processing sites.

D. *The Shannon Entropy is an effective feature to identify Drosha processing sites*

TABLE III. The performance of HairpinSVM and DroshaSVM

Classifiers	SN	SP	MCC	ACC	P
HairpinSVM	0.867	0.988	0.882	0.967	0.938
DroshaSVM	0.908	0.999	0.944	0.998	0.983

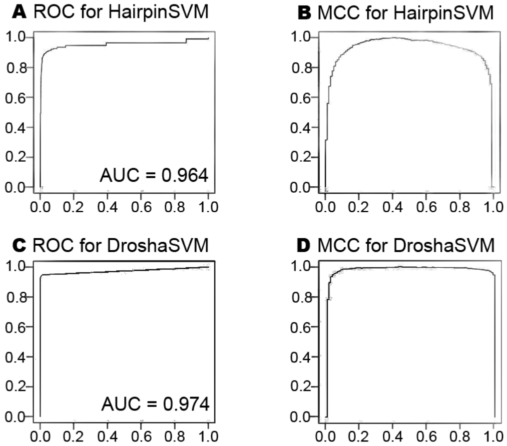

Figure 3. The performance of HairpinSVM and DroshaSVM: (A) The ROC curve for HairpinSVM with the AUC = 0.964. (B) The MCC with the valve curve of HairpinSVM. (C) The ROC curve for DroshaSVM with the AUC = 0.974. (D) The MCC with the valve curve of DroshaSVM.

To our knowledge, Shannon entropy is for the first time to be used in Drosha processing site prediction. The Shannon entropy was proved to be an effective measurement in ncRNA folding [10]. We found that it is an important feature in Drosha processing site prediction (Fig. 4B). The Shannon entropy got higher F-scores than base pair information. After the removal of the Shannon entropy, the AUC of ROC curve for DroshaSVM decreased from 0.974 to 0.886.

We surveyed the scores of the true sites and sites within 3nt calculated by DroshaSVM with or without using the Shannon entropy. We can see from Fig. 5 that the Shannon entropy can greatly reveal the differences between the true and pseudo Drosha processing sites nearby. The average score of true sites is about 0.739, while the average score of pseudo sites close to the true ones is less than 0.006. After removing the Shannon entropy, the average score of true

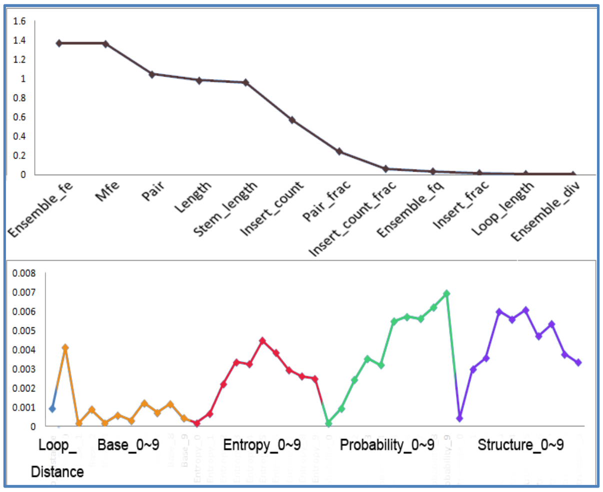

Figure 4. The F-score of features used in HairpinSVM and DroshaSVM. (A) The F-score for HairpinSVM, sort by the value of F-score. (B) The F-score for DroshaSVM, different feature classes are marked with different colors.

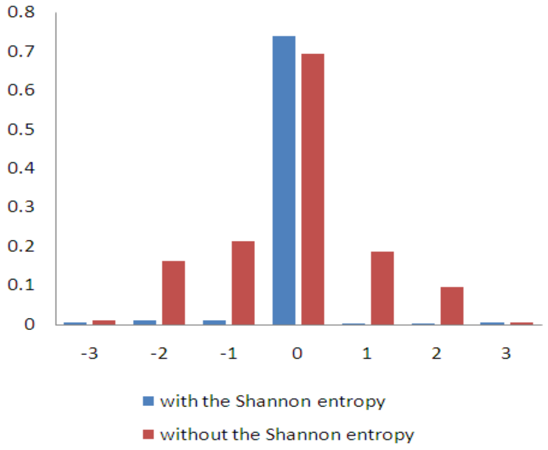

Figure 5. The average score of true Drosha processing sites and neighboring sites given by DroshaSVM with and without the Shannon entropy.

sites is slightly decreased to 0.694, while the average score of position was higher than 0.112. These results demonstrated that the Shannon Entropy is an effective feature to identify Drosha processing sites.

IV. Discussion and Conclusion

Accurately identifying Drosha processing sites is a critical step for miRNA identifying and understanding of miRNA maturation. We presented a Drosha processing site classifier termed DroshaPSP with high prediction accuracy by integrating the sequence and structure information with a two-step prediction model.

The Shannon Entropy is a novel dynamical feature used to predict Drosha processing sites, which is helpful to clearly classify the true processing sites from the nearby pseudo sites. In the previous research of Drosha processing site prediction, it is a noticeable problem that the true sites and the sites within 2nt were scored similarly by their Microprocessor SVM. So we were interested in mining features that can distinguish Drosha processing sites and the neighboring sites sufficiently. After integrating the Shannon entropy feature, we found that DroshaPSP exhibits more powerful capability in distinguishing the true sites from the pseudo sites nearby.

We failed to compare the DroshaPSP with Microprocessor SVM, in which the parameters were trained for Drosha processing sits of human miRNAs. In previous studies, researchers have revealed several differences between human and Drosophila melanogaster miRNAs. For instances, the cleavage partners of Drosha in human and Drosophila are different. Thus, directly comparing two prediction models trained with miRNAs from these two species may lead to the unfair results.

In the future, we will extensively evaluate the performance of DroshaPSP with the prediction model trained on Drosha processing sites from other species including human. In addition, we are planning to develop a stand-alone implement with parallel computation option for Drosha processing site recognition on different operation systems.

Acknowledgment

This study was supported by the National Natural Science Foundation of China (90608020 and 30971642), the Program for New Century Excellent Talents in University (NCET-060651), the National Platform Project of China (2005DKA64001), the Ministry of Education of China (20050487037) and the Natural Science Foundation of Hubei Province of China (2009CDA161).

References

[1] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function (Reprinted from Cell, vol 116, pg 281-297, 2004)," Cell, vol. 131, pp. 11-29, 2007.

[2] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, et al., "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs," Nature, vol. 433, pp. 769-773, 2005.

[3] K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai, "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila," Cell, vol. 130, pp. 89-100, Jul 13 2007.

[4] J. J. Han, Y. Lee, K. H. Yeom, Y. K. Kim, H. Jin, and V. N. Kim, "The Drosha-DGCR8 complex in primary microRNA processing," Genes & Development, vol. 18, pp. 3016-3027, Dec 2004.

[5] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, et al., "The nuclear RNase III Drosha initiates microRNA processing," Nature, vol. 425, pp. 415-419, 2003.

[6] Y. Feng, X. Zhang, Q. Song, T. Li, and Y. Zeng, "Drosha processing controls the specificity and efficiency of global microRNA expression," Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, vol. 1809, pp. 700-707, 2011.

[7] S. Kadener, J. Rodriguez, K. C. Abruzzi, Y. L. Khodor, K. Sugino, M. T. Marr, 2nd, et al., "Genome-wide identification of targets of the drosha-pasha/DGCR8 complex," RNA, vol. 15, pp. 537-45, Apr 2009.

[8] S. A. Helvik, O. Snove, Jr., and P. Saetrom, "Reliable prediction of Drosha processing sites improves microRNA gene prediction," Bioinformatics, vol. 23, pp. 142-9, Jan 15 2007.

[9] M. Huynen, R. Gutell, and D. Konings, "Assessing the reliability of RNA folding using statistical mechanics," J Mol Biol, vol. 267, pp. 1104-12, Apr 18 1997.

[10] E. Freyhult, P. P. Gardner, and V. Moulton, "A comparison of RNA folding measures," BMC Bioinformatics, vol. 6, p. 241, 2005.

[11] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," Nucleic Acids Research, vol. 36, pp. D154-D158, 2008.

[12] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, et al., "The Ensembl genome database project," Nucleic Acids Research, vol. 30, pp. 38-41, Jan 1 2002.

[13] C. J. C. Burges, "A tutorial on Support Vector Machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, Jun 1998.

[14] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, p. 27, 2011.

[15] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," Bioinformatics, vol. 16, pp. 412-424, May 2000.

[16] Y.-W. Chen and C.-J. Lin, "Combining SVMs with Various Feature Selection Strategies

[17] Feature Extraction." vol. 207, I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, Eds., ed: Springer Berlin / Heidelberg, 2006, pp. 315-324.