

Support Vector Machines for Predicting microRNA Hairpins

Karol Szafranski¹, Molly Megraw^{1,2,4}, Martin Reczko⁵, and Artemis G. Hatzigeorgiou^{1,2,3}

¹Center for Bioinformatics, ²Department of Genetics, School of Medicine, ³Department of Computer and Information Science, School of Engineering, ⁴Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA, and

⁵Institute of Computer Science, Foundation of Research and Technology, Hellas, Heraklion, Greece

Abstract - *microRNAs (miRNAs) are 20-22 nt noncoding RNAs which are rapidly emerging as crucial regulators of gene expression in plants and animals. Identification of the hairpins which yield mature miRNAs is the first and most challenging step in miRNA gene prediction. We believe this step can best be achieved with biologically motivated feature design and classification techniques which account for the dependencies inherent in any set of hairpin features. We present DIANA-microH, a tool for predicting microRNA hairpins with high specificity and sensitivity. DIANA-microH implements a Support Vector Machine classifier trained on a set of structural and evolutionary features characteristic of miRNA hairpins. DIANA-microH introduces a unique structural feature motivated by a consideration of how enzymatic cleavage occurs. On test data, the SVM classifier achieved an accuracy of 98.6%. DIANA-microH is applied to chromosome 21 to provide a set of highly probable miRNA hairpins for future laboratory testing.*

Keywords: microRNA, SVM, gene prediction

1 Introduction

A new paradigm of gene expression regulation has emerged recently with the discovery of microRNAs (miRNAs). They are commonly 20-22 nucleotides (nt) long and derive from ~90 nt RNA hairpin structures. miRNAs usually hybridize with imperfect complementarity to the 3' untranslated regions (UTR) of mRNA targets. It has been found that they direct degradation or translational repression of their mRNA targets. The immense potential of small RNAs as controllers of gene networks is just beginning to unfold.

The biogenesis of miRNAs begins with the transcription of a primary transcript (pri-miRNA) which can be up to several kilobases in length [1]. A portion of this primary transcript forms a hairpin structure, which is recognized and cleaved by the enzyme Drosha into a miRNA precursor (pre-miRNA) [2]. The end of the stem portion of a miRNA precursor hairpin defines one end of the double-stranded RNA duplex which will yield the mature miRNA. The pre-miRNA hairpin is exported from the nucleus to the cytoplasm via nuclear export factor Exportin 5 and Ran-GTP cofactor [3-5]. In the cytoplasm,

the RNase enzyme Dicer cleaves the stem of the pre-miRNA to produce a double-stranded RNA duplex containing the mature miRNA sequence on one strand [6-8]. This duplex is unwound and the mature miRNA is incorporated into the RNA-induced silencing complex (RISC) for transport to its target.

Around 200 miRNAs have been identified in human. The total number of miRNAs in each organism is unknown but is estimated to represent ~1% of all genes [9]. Almost all known human miRNAs are (~100%) conserved in the mouse [10], and one third of *C. elegans* miRNAs have vertebrate homologs [9]. Quantitative analysis of miRNA levels in *C. elegans* and HeLa cells shows that overall, miRNA molecules are fairly abundant, ranging from 1,000 to 100,000 copies per miRNA per cell.

The first algorithms for the computational prediction of miRNA genes have been published only recently. The algorithm miRscan was applied to the human and *C. elegans* genomes [9, 11]. Potential miRNA stem-loops are located by sliding a 110-nt window along both strands of the genome. The sequence in the window is folded with the secondary structure prediction program RNAfold [12]. MiRscan further evaluates conserved stem-loops by passing a 21-nt window along each stem-loop and assigning a log-likelihood score to each window. This score provides a measure used to compare the degree to which its attributes resemble those of the experimentally verified mature miRNAs. The determination is based on seven features that characterize the miRNA candidate, examples of these are: the location of the miRNA on the hairpin, the number of base pairs in the 21-nt candidate miRNA and the conservation of the 5' portion of the window. Another approach for prediction of miRNA genes in *C. elegans* was published in [13]. This approach places the emphasis on the conservation of predicted miRNAs across three species. In [14], the evolutionary divergence of different parts of the precursor is closely investigated. Predicted stem loops must have less divergence in both stems compared to the loop, and the position of the miRNA candidate should have perfect conservation. Finally, in the most recent investigation, the prediction performance in *C. elegans* is improved through the scoring of patterns upstream/downstream of the predicted genes and the occurrence of a conserved motif 200 bp upstream from the

stem loop [15]. The estimation of the full number of miRNA genes from these studies is around 100-300 for *C. elegans* [13], around 110 for *Drosophila* [14] and around 250 for human [9].

When considering the number of miRNAs yet to be identified in these species, it is important to note that gene numbers estimated by these programs (MirScan, miRseeker) rest on the assumption that the stem loops of the rare, difficult-to-clone miRNAs will show patterns of conservation and pairing resembling those of the abundant, easily-cloned miRNAs [16]. Also, more recently identified mammalian miRNA genes appear relatively less likely to be conserved in organisms such as fish [16]. If this is the case, the estimates of the current number of miRNA genes in the genome will be too low.

For this reason we believe that there is a strong need for carefully designed improvements and innovative new methods to predict miRNA genes. Identification of the hairpins which yield mature miRNAs is the first and most challenging step in miRNA gene prediction, and we present DIANA-microH for this purpose. The goal of DIANA-microH is to predict miRNA hairpins with high specificity and sensitivity. DIANA-microH implements a Support Vector Machine (SVM) classifier trained on a set of structural and evolutionary features characteristic of miRNA hairpins. DIANA-microH introduces a unique structural feature motivated by a consideration of how enzymatic cleavage occurs. An SVM classifier is a machine learning technique well-suited to handle the dependencies inherent in any set of hairpin features.

2 Methods

2.1 Data Sets

2.1.1 Sequence data sets for SVM training.

Positive training data were the human miRNA hairpins listed in the RFAM database [17], along with their genomic coordinates referring to human genome build NCBI34. We used final exons of protein-coding genes, according to UCSC RefSeq annotation, as prediction templates for generating negative training data. We used only final exons that contained at least 100 bp of 3' UTR sequence, that were not involved in alternative polyadenylation or splicing, and that had a maximum total length of at most 300 bp. Only those hairpins that were fully contained within a UCSC AXT "net" fragment [18] were included into the data sets.

2.1.2 Partitioning of data sets.

Machine learning algorithms such as SVM classifiers require carefully chosen training, validation, and test data sets in order to maximize performance in the training and

appropriately evaluate performance in the testing. We define these datasets by partitioning the human miRNA hairpins described above into training/validation and test sets using a technique which avoids overrepresentation by similar miRNA precursors, while ensuring that no family of mature miRNAs is split between these sets. Specifically, the full set of human miRNA precursors is clustered into groups by precursor similarity, and the full set of mature miRNAs resulting from these hairpins is independently clustered into miRNA families. Then the set of precursor data is divided into training/validation and test sets in proportions as close as possible to 75%/25% such that only one hairpin per hairpin group will be placed into any output subset and such that no mature miRNA group will be split among these sets. In this way, similar hairpins from a particularly abundant group cannot skew the training process. We also insure that training/validation and test sets do not contain closely related hairpins. The SVM training procedure used the training/validation set to train model parameters with 5-fold cross validation.

2.1.3 Chromosome 21 evaluation sequence.

The human chromosome 21 sequence contained within UCSC AXT "net" human-mouse alignments, referring to genome freezes hg16/mm4 (NCBI builds 34 and 32, respectively), was the sequence used to perform an evaluation run of DIANA-microH. These sequences represent the 35% most conserved fraction of the chromosome that shows local similarity to mouse sequence.

2.2 Algorithm

2.2.1 Features of the miRNA hairpins.

First the secondary structure of the RNA sequence under investigation must be calculated. For two adjacent sequence windows of 52 nt each we calculated the hairpin stem structure that has minimum free energy. The energy minimization algorithm [19] is a modification of the Needleman-Wunsch algorithm. This approach and our implementation is described in detail in the supplementary material of [20]. Energy tables for canonical (Watson-Crick) and G-U wobble dinucleotide base pairs as well as the destabilizing contribution of interior loops were reported by the Turner group [21]. We considered only hairpins having a total free energy below -25 kcal/mol at 37 °C.

From this secondary structure we can then determine the characteristic features for the hairpin. The features used by DIANA-microH which have appeared in other miRNA prediction tools are:

- (1) *Free Energy* is the total free energy resulting from the primary hairpin structure prediction.

(2) *Paired Bases* is the number of nucleotides predicted to be in a hydrogen-bonded state.

(3) *Loop Length* is the length of the hairpin loop of the predicted secondary structure.

(4) *Arm Conservation* is the average sequence identity throughout the sequence of the stem substructure. We first identify a fragment in the UCSC AXT resource of local pairwise alignments (either human-mouse or human-rat, “tight” or “net”) that fully contains the predicted hairpin. The hairpin is mapped to the alignment and the sequence identity count is derived from the alignment window that corresponds to the “arms” of the structure, i.e. both strands of the stem portion.

DIANA-microH introduces the following two new features for the purpose of detecting microRNA hairpins:

(5) *GC Content* is defined as the fraction of G and C nucleotides in the structure prediction window.

(6) *Stem Linearity* is defined as the largest possible section of the stem subregion that is likely to form a mostly linear double-stranded conformation. The underlying algorithm first considers every base-paired stem position to be a potential start of such a linear section. The method iterates over either a series of paired bases or loop-forming subsequences. Progression through a series of paired bases causes an update of the direction of a helix direction vector, as does progression through a symmetric interior loop. By contrast, progression through an asymmetric interior loop gives rise to the generation or addition of a deviation vector, which has a fixed length and the same direction as the current helix direction vector. The length of this deviation vector is added to the cumulative deviation score for every subsequent duplex progression that is made. Iteration ends if this cumulated linearity deviation score breaks a certain threshold. In this way, this score encodes the “bendedness” of the segment under examination.

2.2.2 Classifier.

DIANA-microH uses Support Vector Machines (SVMs) [22], which have the significant advantage that they can recognize dependencies among the given features. Other machine learning and statistical techniques typically assume complete independence among the features, an unrealistic assumption for features on miRNA hairpins such as free energy, number of loops and bulges, and base-pairing relationships.

To understand how an SVM classifier works, first think of the task of separating two classes of points in space. If two classes are linearly separable, we can define optimal separating hyperplanes for their separation. If the

classes overlap, as is the case with miRNA gene features, we need to produce nonlinear boundaries to separate them. SVMs allow for this by constructing a linear boundary in a large, transformed version of the feature space.

Here we use the C-Support Vector Classification (C-SVC) [23, 24] algorithm. Given training vectors $\mathbf{X}_i \in \mathfrak{R}^n$, $i = 1, \dots, K, \dots, l$ and for each vector corresponding class assignments $y^i \in \{1, -1\}$, C-SVC solves the following optimization problem:

$$\begin{array}{ll} \min & \frac{1}{2} \omega^T \omega + C \sum \xi_i \\ \text{Subject to} & y_i (\omega^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i > 0 \end{array}$$

The training vectors \mathbf{X}_i are mapped into a higher dimensional space by the function ϕ . Then the SVM finds a linearly separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term and $K(\mathbf{X}_i, \mathbf{X}_j) \equiv \phi(\mathbf{X}_i)^T \phi(\mathbf{X}_j)$ is called the kernel function.

We use the RBF kernel defined by $K(\mathbf{X}_i, \mathbf{X}_j) \equiv \exp(-\gamma \|\mathbf{X}_i - \mathbf{X}_j\|^2)$, $\gamma > 0$ which nonlinearly maps samples into a higher dimensional space. [25] shows that the linear kernel with a penalty parameter C has the same performance as an RBF kernel with parameters (C, γ) , so the linear kernel is a special case of RBF. Additionally, the sigmoid kernel behaves like an RBF for certain parameters [26]. Compared with the polynomial kernel, there are also less hyperparameters required for an RBF kernel to be optimized.

The SVM implementation used by DIANA-microH is the C program package libsvm-2.6 [27]. Training and testing was performed on feature data that was scaled to a value range of [-1.0,1.0]. Scripts in Perl, bash and awk were used to wrap the SVM in order to incorporate it into the prediction pipeline.

3 Results

3.1 Discriminative Power of Features

A powerful aspect of using an SVM classifier is that it allows one to focus on the design of the features. The features must ultimately describe how an enzyme recognizes a miRNA hairpin for cleavage, since in essence we are trying to mimic the natural process of distinguishing the “right” hairpin. This is a fundamental part of how DIANA-microH is able to achieve its accuracy. Here we describe the extent to which each feature was able to partition real miRNA hairpins from negative controls.

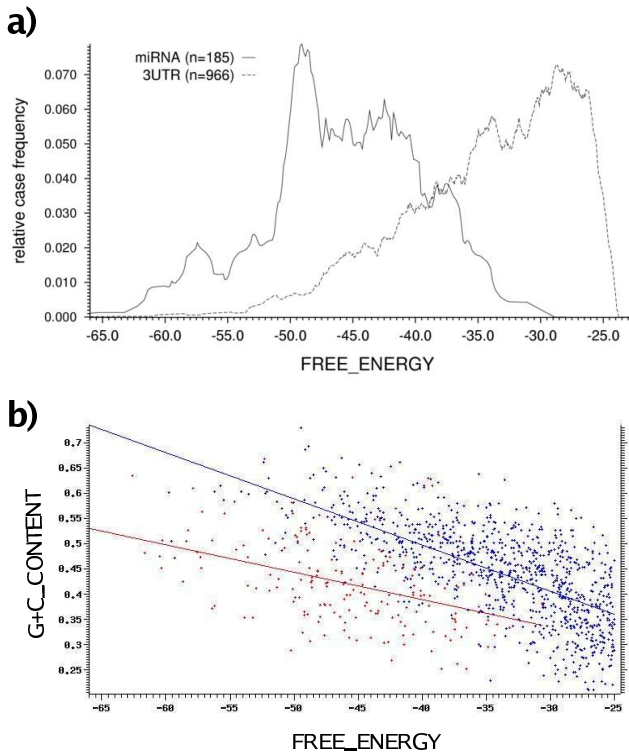


Fig. 1. Discriminatory power of total free energy, either alone or in conjunction with a consideration of the GC content of the sequence. **a)** Plot of the relative case frequencies of hairpins as a function of their total free energy, for both true miRNAs (solid line) and 3'UTR controls (dotted line). **b)** Case correlations of total free energy of the secondary structure and the GC content of its underlying sequence, for both true miRNAs (red) and 3'UTR controls (blue).

Free Energy, Paired Bases, GC Content. Within the hairpin prediction window, miRNA hairpins exhibit rather long, near-perfect double stranded regions. Two measures, the total free energy of the predicted fold and the number of paired bases, have been implemented in our prediction tool to account for this feature. The hairpins used as negative controls less frequently exhibit low total free energy values or a high fraction of paired bases. In general, the free energy value is correlated with the GC content of the sequence, for both miRNA hairpins and controls (see Figure 1b). By exploiting this observation, we were able to achieve a far better separation between real miRNA hairpins having particularly high free energy values (those hairpins in the right tail of the miRNA hairpin energy distribution in Figure 1a) and controls having particularly low free energy values (those controls in the left tail of the negative control energy distribution in Figure 1a): miRNA hairpins in this ambiguous energy value range will have lower GC content than control structures with similar free energy values (see Figure 1b).

Loop Length. Hairpin loops require a length of 10 nt or longer to be enzymatically processed into miRNAs, as has been shown in [28]. The same authors and others [29] have noted that secondary structure predictions based on energy minimization often result in miRNA hairpin loops smaller than 10 nt. It has been suggested that Drosha, the enzyme responsible for first-step processing of the miRNA primary transcript, melts weakly paired bases inside the loop prior to processing. We observe that the predicted structures of miRNA transcripts have a frequency peak at hairpin loop lengths of 12 to 14 nt, while controls usually have shorter loop lengths (see Figure 2).

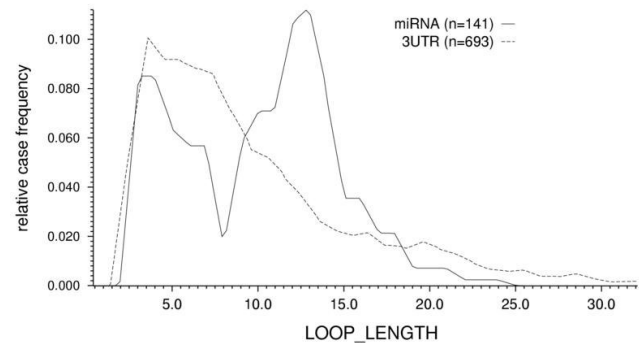


Fig. 2. Discriminatory power of hairpin loop length. Shown is the plot of the relative case frequencies of hairpins as a function of the length value, for both true miRNAs (solid line) and 3'UTR controls (dotted line).

Arm Conservation. Evolutionary sequence conservation is a prominent characteristic of miRNA genes. The 20% most conserved fraction of the human genome (i.e. UCSC's AXT "net" human-mouse alignments) contains all reported human miRNAs. The most weakly conserved miRNA, hsa-mir-197, has 49% sequence identity between human and mouse. We used the UCSC AXT resource of pairwise local genomic alignments for retrieving sequence identity values for human versus rodent (mouse or rat). Other authors have derived refined conservation profiles from pairwise sequence alignments [14] and we also experimented with similar profiles. However, we chose to incorporate sequence conservation properties in the form of mean sequence identity for the predicted stem substructure, obtaining a sensitive (and discriminative) measure while avoiding over-training and keeping the variance of the property measure low. Mean sequence identity outperformed schemes which used separate conservation profiles for the loop and stem arms.

Stem Linearity. Another criterion for scoring a predicted miRNA hairpin is to identify the largest possible section of the stem that is likely to form a highly linear

double-stranded conformation. This feature was biologically motivated by the observation that non-miRNA hairpin structures predominantly contain bulges and asymmetric interior loops that will result in irregular bending of the stem substructure while true miRNAs are free of those loops, or the loops occur in a sequential pattern where local bends are likely to compensate each other resulting in an overall near-linear double stranded conformation. The discriminative power of this feature supports our hypothesis that the enzyme Drosha recognizes and cleaves hairpins most effectively when the stem portion of the hairpin is linear (see Figure 3).

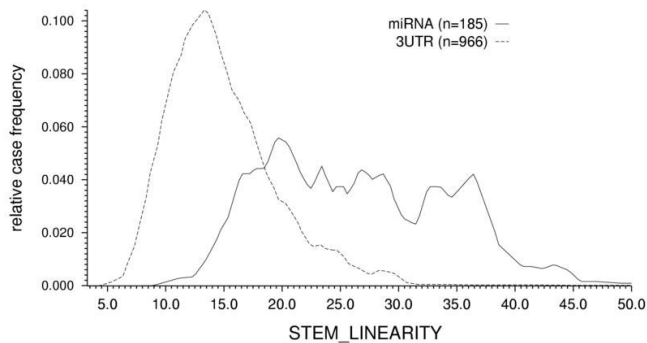


Fig. 3. Discriminatory power of feature Stem Linearity. Shown is the plot of the relative case frequencies of hairpins as a function of the stem linearity feature value, for both true miRNAs (solid line) and 3'UTR controls (dotted line).

3.2 SVM Performance

The performance of our algorithm was tested on a separate set containing 45 positive hairpins (true miRNAs) and 243 negative hairpins (predicted from the 3' UTR sequence). On this test set, the SVM classifier performed with an accuracy of 98.6% (284 out of 288 correct classifications, 2 false positives, 2 false negatives). We did not adjust the classification boundary in order to include all true miRNAs into the positive class since this would have resulted in an unacceptably low specificity.

3.3 Evaluation on Chromosome 21

Human chromosome 21 was used to perform an evaluation run of DIANA-microH. This 12 Mbp of sequence represents about 35% of the euchromatic portion of that chromosome. Secondary structure prediction and SVM scoring was done as outlined. The whole prediction process for both strands took 33 hours on a single Pentium Xeon 2.6 GHz processor, corresponding to 3 hours per Mbp of double-stranded sequence. It turned out that the first step of secondary structure prediction represents the

computational bottleneck of the entire prediction pipeline. However, the parameters could be re-tuned. For example, a less conservative choice for the interior loop size limit L , currently set to 11 nt, would clearly reduce computational time. Moreover, DIANA-microH has an architecture that allows for parallelization of the first step of prediction which includes the secondary structure prediction and determination of the feature values.

The prediction yielded 35 hairpins with outstandingly high SVM scores. This group contained all four miRNA hairpins that are listed in RFAM on Chromosome 21 (hsa-let-7c, hsa-mir-99a, hsa-mir-125b-2, hsa-mir-155). One quite encouraging aspect of the top hit list is that it does not contain predictions for the negative strand of known miRNAs. Obviously, DIANA-microH predictions are highly strand-specific although we have not given any special attention to this aspect during the training process.

Since many hairpins in the group of top hits were made up of microsatellites and minisatellites, e.g. $(GT)_n$, the results were post-filtered using multi-phase nucleotide correlation recording (unpublished software) and applying a repetitiveness threshold that would successfully retain all known microRNAs. This satellite filter removed 24 hairpin predictions, leaving 15 novel cases (see Table 1). This list may be further narrowed, considering the location of the hairpins in relation to surrounding or overlapping genes. There are some hairpins predicted to reside in the coding sequence (CDS) of known protein-coding genes, and we are skeptical that these cases represent miRNAs since there are no known cases of CDS overlap with a miRNA gene. It is surely possible to filter such predictions automatically. Moderate to high GC content and the typically high degree of evolutionary conservation is likely to give rise to an increased rate of false positive predictions for CDS regions. However, it is interesting to note that in Table 1 all hairpins predicted in CDS regions reside in the gene SON. The coding sequence of this gene is particularly conducive to forming hairpins with features very much like those observed in miRNA hairpins. SON is extensively alternatively spliced, so it may be the case that these hairpins reside in the intronic structures of some SON splice forms. Provocatively, hairpins have been shown to be involved in splice regulation [30].

Table 1. Top-scoring microRNA hairpins predicted by DIANA-microH on chromosome 21.

#	hg16 Coordinates	SVM Score	Genetic Region	Genetic Locus	Notes
1	chr21: 33844136- 33844240	0.99982	CDS	SON	
2	chr21: 33843623- 33843727	0.99903	CDS	SON	
3	chr21: 18973751- 18973855	0.99893	intron	cDNA support	close to #9
4	chr21: 16834011- 16834115	0.99882	intron	C21orf34	known: hsa-let-7c
5	chr21: 25868146- 25868250	0.99879	intergenic		known: hsa-mir-155
6	chr21: 33843086- 33843190	0.99830	CDS	SON	
7	chr21: 16833267- 16833371	0.99816	intron	C21orf34	known: hsa-mir-99a
8	chr21: 33852618- 33852722	0.99806	CDS alternat.	SON	
9	chr21: 18973695- 18973799	0.99723	intron	cDNA support	close to #3
10	chr21: 33845352- 33845456	0.99689	CDS	SON	
11	chr21: 27556982- 27557086	0.99620	intergenic		
12	chr21: 33852619- 33852723	0.99358	CDS alternat.	SON	
13	chr21: 38579832- 38579936	0.99350	intron	KCNJ15	Overlaps MER53 repeat
14	chr21: 31655002- 31655106	0.98817	intron	TIAM1	
15	chr21: 33844523- 33844627	0.98776	CDS	SON	
16	chr21: 42590952- 42591056	0.98440	intron	ABCG1	
17	chr21: 16884420- 16884524	0.98357	intron	C21orf34	known: hsa-mir-125b-2
18	chr21: 32455864- 32455968	0.98093	intergenic		
19	chr21: 23787536- 23787640	0.98058	intergenic		

4 Discussion

The application of DIANA-microH to chromosome 21 provides 7 high-scoring candidates which we believe are appropriate for laboratory testing: miRNA hairpins numbered 3, 9, 11, 14, 16, 18, and 19 in Table 1. These hairpins are located in the introns of protein-coding genes and in intergenic regions, and do not overlap repeat regions.

Ultimately the prediction of mature miRNA sequences will be a continuation of this work. Other miRNA prediction tools use similar core features for the prediction of hairpins and mature miRNAs [9, 11, 13, 14]. The most recent developments tend to move in the direction of creating more detailed versions of the existing core features; for example, creating more detailed conservation profiles. The goal of DIANA-microH, however, was to return to biological basics to add a few features which provide more specificity while avoiding loss of sensitivity. One feature, *Stem Linearity*, relates to the biological process of substrate recognition and therefore yields strong discriminatory power. The power of the feature set for DIANA-microH was harnessed by an SVM classifier which successfully separated real miRNA hairpins from negative controls with an accuracy of 98.6%.

5 Acknowledgements

We thank Petko Fitziev for his help in scripting during the initial development stage of our miRNA secondary structure prediction code. We also thank Andrei Kouranov and Zissimos Mourelatos for their helpful discussion during the early stages of this work. The work by A. H. and M. M. was supported in part by NSF Career Award DBI-0238295.

6 References

- [1] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "MicroRNA genes are transcribed by RNA polymerase II," *Embo J*, vol. 23, pp. 4051-60, 2004.
- [2] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V. N. Kim, "The nuclear RNase III Drosha initiates microRNA processing," *Nature*, vol. 425, pp. 415-9, 2003.
- [3] R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen, "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs," *Genes Dev*, vol. 17, pp. 3011-6, 2003.
- [4] M. T. Bohnsack, K. Czaplinski, and D. Gorlich, "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs," *Rna*, vol. 10, pp. 185-91, 2004.
- [5] E. Lund, S. Guttinger, A. Calado, J. E. Dahlberg, and U. Kutay, "Nuclear export of microRNA precursors," *Science*, vol. 303, pp. 95-8, 2004.

- [6] A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello, "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing," *Cell*, vol. 106, pp. 23-34, 2001.
- [7] G. Hutvagner, J. McLachlan, A. E. Pasquinelli, E. Balint, T. Tuschl, and P. D. Zamore, "A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA," *Science*, vol. 293, pp. 834-8, 2001.
- [8] R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk, "Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*," *Genes Dev*, vol. 15, pp. 2654-9, 2001.
- [9] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel, "Vertebrate microRNA genes," *Science*, vol. 299, pp. 1540, 2003.
- [10] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of novel genes coding for small expressed RNAs," *Science*, vol. 294, pp. 853-8, 2001.
- [11] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel, "The microRNAs of *Caenorhabditis elegans*," *Genes Dev*, vol. 17, pp. 991-1008, 2003.
- [12] I. L. Hofacker, S. Fontana, W. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte f. Chemie*, vol. 125, pp. 167-188, 1994.
- [13] Y. Grad, J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim, "Computational and experimental identification of *C. elegans* microRNAs," *Mol Cell*, vol. 11, pp. 1253-63, 2003.
- [14] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin, "Computational identification of *Drosophila* microRNA genes," *Genome Biol*, vol. 4, pp. R42, 2003.
- [15] 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.
- [16] D. P. Bartel and C. Z. Chen, "Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs," *Nat Rev Genet*, vol. 5, pp. 396-400, 2004.
- [17] S. Griffiths-Jones, "The microRNA Registry," *Nucleic Acids Res*, vol. 32, pp. D109-11, 2004.
- [18] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database," *Nucleic Acids Res*, vol. 31, pp. 51-4, 2003.
- [19] I. Tinoco, Jr., P. N. Borer, B. Dengler, M. D. Levin, O. C. Uhlenbeck, D. M. Crothers, and J. Bralla, "Improved estimation of secondary structure in ribonucleic acids," *Nat New Biol*, vol. 246, pp. 40-1, 1973.
- [20] 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.
- [21] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J Mol Biol*, vol. 288, pp. 911-40, 1999.
- [22] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag, 2001.
- [23] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.
- [24] V. Vapnik, *Statistical Learning Theory*. New York: John Wiley, 1998.
- [25] S. S. Keerthi and C. J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput*, vol. 15, pp. 1667-89, 2003.
- [26] C.-J. Lin and H.-T. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," Department of Computer Science and Information Engineering, National Taiwan, 2003.
- [27] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.
- [28] Y. Zeng, R. Yi, and B. R. Cullen, "Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha," *Embo J*, vol. 24, pp. 138-48, 2005.
- [29] J. Krol, K. Sobczak, U. Wilczynska, M. Drath, A. Jasinska, D. Kaczynska, and W. J. Krzyzosiak, "Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design," *J Biol Chem*, vol. 279, pp. 42230-9, 2004.
- [30] Z. R. Liu, B. Lagerbauer, R. Luhrmann, and C. W. Smith, "Crosslinking of the U5 snRNP-specific 116-kDa protein to RNA hairpins that block step 2 of splicing," *Rna*, vol. 3, pp. 1207-19, 1997.