SVM-Based miRNA:miRNA* Duplex Prediction

Karathanasis Nestoras, Tsamardinos Ioannis
Department of Biology, Department of Computer Science
University of Crete
Heraklion, Greece

Armen Angelos P., Tsamardinos Ioannis, Karathanasis Nestoras, Poirazi Panayiota
Institute of Computer Science, Institute of Molecular Biology and Biotechnology
Foundation for Research and Technology - Hellas FORTH
Heraklion, Greece

Abstract—**We address the problem of predicting the miRNA:miRNA* duplex stemming from a microRNA (miRNA) hairpin precursor and we present a SVM-based methodology to address it. Predicting the miRNA:miRNA* duplex is a first step towards identifying the mature miRNA, suggesting possible miRNA targets and ultimately, reducing experimentation effort, time, and cost. We measure the error in terms of the absolute difference of the true and predicted location of all of the four ends of the duplex and/or of each end separately. Our mean absolute error over all ends is 1.61 ± 2.24 nts as measured on a hold-out set of 220 miRNA hairpin precursor sequences. In addition, our tool precisely predicts (with 0 nt deviation) the starting position for 57% and 52% of the miRNAs in the 5' and 3' strands of the same dataset, significantly outperforming the state-of-the-art tool *MaturePred* which achieves 18% and 12%, respectively, on the same task. Overall, our method accurately identifies not only the starting nucleotide of novel miRNA:miRNA* duplexes –and thus individual miRNAs- but also their length, while outperforming the current state-of-the-art tool.**

Keywords- miRNA:miRNA*; duplex; microRNA; SVM; Dicer.

I. INTRODUCTION

MicroRNAs are small, ~22 nucleotides long, single-stranded non-coding RNAs that play an important regulatory role in both animals and plants by binding at target sites on messenger RNAs (mRNAs), leading to mRNA cleavage or translational repression [1]. The primary transcripts of microRNA genes are called *primary miRNAs a*nd consist of a stem-loop ("hairpin") structure extended with long single-stranded tails. The tails are detached (in animals) by the *Microprocessor* complex, whose core component is the RNase III enzyme Drosha, leaving a hairpin-shaped, ~60−70 nts long intermediate with a characteristic 3' overhang of ~2 nt, the miRNA *precursor* (*pre-miRNA*). The pre-miRNA is then exported to the cytoplasm, where it is processed by another RNase III termed *Dicer*. Dicer cleaves the pre-miRNA at a certain distance (~22 nt) from the overhang created by the Microprocessor [2], leaving an RNA duplex with 3' overhangs of ~2 nts called *miRNA-miRNA** duplex. For each individual duplex, one (or both) of its strands ends up as the *mature* miRNA and is loaded into a *RISC* (*RNA-induced Silencing Complex*), where it performs its regulatory functions on target mRNA. The other strand, called *miRNA**, is degraded. It may also be the case that both strands of the duplex correspond to a mature miRNA: only one strand becomes the miRNA each time but with similar frequency [1]. It was recently found that some pre-miRNAs (the so-called *mirtrons*) have a similar structure with regular pre-miRNAs, but enter the miRNA pathway without undergoing processing by Drosha, *i.e.* without undergoing the pri-miRNA stage.

Given the importance of miRNAs in gene regulation, several computational approaches have been developed to complement experimental ones. Most of them focus on the discovery of novel miRNA genes or possible mRNA targets of known miRNAs [3], [4]. As part of miRNA gene discovery, these tools predict certain features of miRNAs such as the starting position of the mature miRNA [5-7], the Drosha cleavage site [8] (which coincides with the start of the mature miRNA on a pri-miRNA) or the mature miRNA molecule on hairpin precursors [9], [10], [11] . In all cases, the results are amenable to improvement as performance accuracy with respect to the identification of the exact mature miRNA molecule remains low.

In this paper, we introduce the problem of identifying the miRNA:miRNA* duplex as a first step in identifying the mature miRNA. We adopt this approach because (a) the duplex is a necessary stage of miRNA biogenesis and (b) given the duplex, it is relatively easy to experimentally determine whether both, or which of the two duplex strands results in the mature miRNA(s). In the experiments presented here we use mouse and human hairpins obtained from version 17.0 of miRBase (release 17.0: April 2011) [12] to devise an algorithm for producing candidate miRNA:miRNA* duplexes from hairpin sequences. We subsequently train a Support Vector Machine (SVM) on these duplexes using the known ones as positive examples and the rest as negative. We optimize the input features feeding into the SVM model using cross-validation and report the performance on a hold-out set. The final prediction of our method on a new hairpin is the candidate duplex that maximizes the score of the SVM model. We measure the prediction error in terms of the absolute difference between the true and predicted location of all of the four ends of the duplex as well as in each end separately. We find that our tool significantly outperforms the state of the art tool *MaturePred*, as well as a *Trivial* locator, when assessed on a common blind test set.

II. PRODUCING CANDIDATE DUPLEXES

In a hairpin sequence, the counting of nucleotide positions starts from the 5' end and continues until the 3' end. A hairpin consists of a double-stranded part, the stem and a sequence of unmatched nucleotides that connects the strands of the stem,

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

called the terminal loop. The strand before the terminal loop is called the 5’ arm of the hairpin while the other is called the 3’ arm. The arms are not perfectly complementary but they form small loops and bulges. We denote a case of a match of a nucleotide at position p on the 5’ arm with one at position q on the 3’ arm with M (p, q), p < q.

A miRNA:miRNA* duplex consists of two hairpin subsequences on each of the two arms, called the 5’ strand and the 3’ strand of the duplex. We can define a duplex by the positions of its four ends on the generating hairpin sequence; we name them k55, k53, k35 and k33 corresponding to the 5’strand 5’end, 5’strand 3’end, 3’strand 5’end and 3’strand 3’end positions, respectively. Notice that, because of the way of counting positions, k55 < k53 < k35 < k33. Fig.1 shows an example of a real hairpin (hsa-mir-17) with all the above quantities annotated.

Typical features of the duplex are the overhangs' length and the length of mature sequences. In human and mouse data the overhang length ranges between -11 nts up to +18 nts for 5’ strand and -16 nts up to +12 nts for the 3’ strand with a modal value of +2 nts on both cases, characteristic of RNase III (Drosha, Dicer) cuts. The mature miRNA length ranges between 17 - 26 nts for the 5’ strand and between 18 – 26 nts for the 3’ strand with an overrepresentation of ~22 nts length in both cases.

In order to produce the candidate duplexes, first we calculate the statistical distributions of the overhangs’ and matures’ lengths on the given training set and remove the overhangs’ outlier values (values that are above or below three times the standard deviation from the mean value). Then we utilize two scanning windows, one on each strand. The length of these windows ranges between the minimum and maximum values of the respective matures’ length distribution. We keep only the candidate duplexes whose overhangs’ lengths are represented in the training set’s distribution. In addition, position k53 reaches up to the loop tip which allows the identification of mature miRNAs that extent into the terminal loop as opposed to previous methods [11]. Importantly, applying this methodology during the testing phase allows us to produce the true duplex in all cases tested. However, the possibility of missing the true duplex due to the removal of outliers in the first step cannot be excluded.

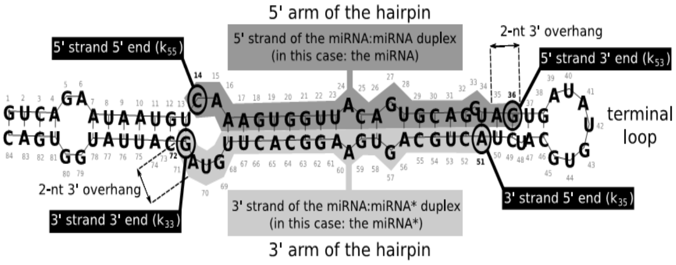
5' arm of the hairpin
5' strand of the miRNA:miRNA duplex
(in this case: the miRNA)
2-nt 3' overhang
5' strand 5' end (k_{55})
5' strand 3' end (k_{53})
terminal
loop
3' strand 5' end (k_{35})
2-nt 3' overhang
3' strand 3' end (k_{33})
3' strand of the miRNA:miRNA* duplex
(in this case: the miRNA*)
3' arm of the hairpin
14
36
51
72

Figure 1. Anatomy of the hsa-mir-17 hairpin

III. DUPLEX VECTOR REPRESENTATION

To select among all candidate duplexes we use a polynomial kernel SVM. To train these models, a fixed-length numerical vector representation of training examples is required. Only nucleotide sequence information is used in the representation. Nucleotide sequences are converted to binary vectors, using a 1-of-4 encoding: bases A, T, G and U are represented as 1000, 0100, 0010 and 0001 respectively. A problem with using a fixed-vector representation is that the strand sequences are of variable size. One solution to this problem is to identify the maximum possible strand length and pad with zeros or some other special value at the end for the missing nucleotides. Unfortunately, this means that the suffix of the nucleotide sequences will be represented with different variables each time. However, it has been demonstrated that end structure and nucleotide sequence are the primary determinants of Dicer specificity when it processes double-stranded or short hairpin RNA [2]; we thus preferred a representation where it is the ends of the sequences that always correspond to the same variables . To do so we pad with zeros in the middle of a sequence, so that the first and the last nucleotides are always represented with the first and last variables respectively. Padding vectors with zeros is common in signal processing to enable signals to reach a prespecified length. While there may be better ways to treat missing information, zero padding does not affect the estimation of performance of the resulting models or invalidates any results

Prior work suggests that the flanking regions around the processing sites of Drosha and Dicer are important for the identification of these cut sites. For example, in [8], maximum performance is achieved when including the region before the pre-miRNA in the representation of a candidate Drosha processing site, while in [10] the same is true when including the regions before and after each candidate miRNA. Accordingly, we include the flanking regions at both ends of each duplex strand in the representation of a candidate duplex. As suggested by earlier work in our lab [10], the exact length of each flanking region was optimized within the set of {12 - 14} nts (see Section IV); the best value found was 14 nts. If a flanking region extends beyond the arm’s boundaries, zero padding at the beginning (for 5’ end flanking regions) or at the end (for 3’ end flanking regions) of the sequence takes place. The complete representation of a candidate duplex (Ds) consists of the encoded nucleotide sequences of both its strands and their flanking regions. In Fig.2 the various parts consisting the vector representation of the true miRNA:miRNA* duplex stemming from the hsa-mir-17 hairpin are shown.

IV. PRODUCING THE SVM MODEL

We used version 17.0 of miRBase (release 17.0: April 2011) [12] which contains 16772 entries representing hairpin precursor miRNAs, expressing 19724 mature miRNA products, in 153 species. These hairpin precursor sequences do not always correspond to the exact pre-miRNA sequence but consist of the latter extended with some flanking nucleotides. We selected only human and mouse sequences and first filtered out all entries with unknown duplexes and/or multi-branch structures. Out of the remaining hairpins, (about 2% of

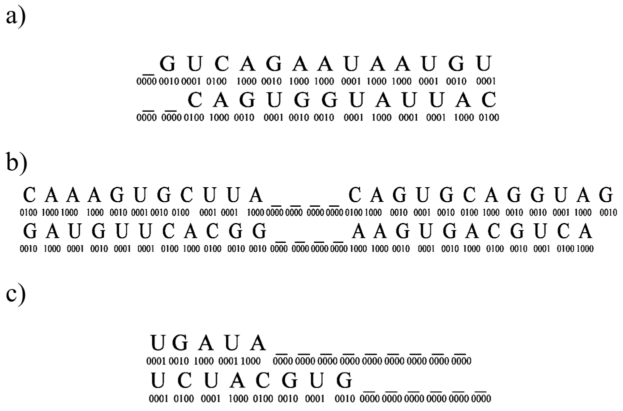
a)
_ G U C A G A A U A A U G U
_ _ C A G U G G U A U U A C
b)
C A A A G U G C U U A _ _ _ _ C A G U G C A G G U A G
G A U G U U C A C G G _ _ _ _ A A G U G A C G U C A
c)
U G A U A _ _ _ _ _ _ _ _ _
U C U A C G U G _ _ _ _ _ _

Figure 2. Vector representation for the 5’ strand of the true miRNA:miRNA* duplex stemming from the has-mir-17; the same methodology is also used for the 3’strand. a) 5′strand 5′end (top) and 3’strand 3’ end (bottom) 14-nts flanking region nucleotide sequence, b) 5′ strand (top)-3’ strand (bottom), duplex nucleotide sequence. Zero padding takes place in the middle, in order to reach 27 nts in the 5’ strand and 26 nts in the 3’ strand which are the maximum values found in the training set. c) 5′strand 3′end (top), 3’strand 5’ end (bottom) 14-nts flanking region nucleotide sequence. Zero padding takes place at the end, because the 14-nts flanking region extends beyond the 5′arm 3′end/ 3’arm 5’ end , which is defined to be halfway from the terminal loop.

available human and mouse sequences were filtered out due to non-hairpin structures or database inconsistencies) we selected 656 human - mouse hairpin sequences and used a 5-fold cross-validation technique to train our model on these sequences.

Specifically, for each miRNA hairpin in the training set, we generated all possible duplexes as detailed in section II, resulting in ~10.000 candidates per hairpin. Duplexes containing experimentally verified miRNA strands were labeled positive and the rest were labeled as negative. Only 100 randomly selected negatives duplexes per positive sample were used for training to reduce training time.

Training samples were used both for optimizing the flanking regions and for learning a final SVM model. The SVM software used was the MATLAB interface for LIBSVM (version 2.88) [13]. The kernel of the SVM used was the full polynomial K (x_i, x_j) = (x_i · x_j + 1)^d, where · represents the inner product of the vectors and d is the degree of the polynomial. The distribution of the two classes is quite unbalanced (1:100); we thus used two cost parameters, one for each class. Specifically, the cost parameters for the positive and negative class were (number of samples)/ (2*number of positive samples) and (number of samples)/ (2*number of negative samples), respectively.

The degree of the kernel within the set {1, 2, 3, 4} and the size of each flanking region within {0, 4, 8, 12, 13, 14} nts had been optimized in a previous version of the algorithm. Both of them were selected based on intuition and preliminary anecdotal experiments; *the test set was never employed in the latter and so the final performance estimation remains unbiased*. The best performing combination of parameters was selected using 5-fold cross-validation: degree 3 of the kernel and length 14 for the flanking regions. The flanking region within {12, 13, 14} nts, was optimized again, using the previously described data set, and was again found to be 14 nts. The flanking search space was restricted to that range due to time limitations. The final performance on the test set and the performance during cross-validation are measured in terms of predicting the exact location of the duplex, which is informative and intuitive to a biologist.

For comparison reasons, a final SVM model using the above parameters was re-trained on a slightly different training set: Specifically, we downloaded the positive dataset of *MaturePred*’s animal model from its website [9]. The positive dataset contained the experimentally verified animal pre-miRNAs of mirBase 14 which consists of 4419 entries representing hairpin precursor miRNAs, from 91 species. Again, entries with unknown duplexes and/or multi-branch structures were filtered out and the remaining 1784 hairpins were used to train the SVM model. This was done in order to generate an SVM model that is directly comparable to *MaturePred*, as it was trained on a subset – only on the hairpins known to contain a duplex – of *MaturePred*’s training set.

V. PRODUCING A BASE LINE TRIVIAL MODEL

We also compare our tool to a *Trivial* classifier, whereby each of the four ends of the duplex is represented by its average location in the training set (this is the same training set used to produce our final SVM model). Specifically, for any given hairpin, the location of each of the four ends in known duplexes is found by calculating its distance from the tip of the terminal loop. This is done over all hairpins in the training set and the average distances are then used to generate the predictions of the ‘*Trivial*’ classifier for any new hairpin in the hold out set. The terminal loop tip was chosen as the reference point as it does not depend on the length of the pre-miRNA flanking regions included in the hairpin sequence.

VI. RESULTS

To evaluate the training performance of our tool we select 656 human mouse hairpins and use a 5-fold cross validation as described in section IV. The mean absolute error of *DuplexSVM* taken over the four ends of the predicted duplex for all samples in the training set is 1.57 ± 1.96 nts and the respective values for the *Trivial* classifier are 1.83 ±1.57 nts. Table I shows the percentage of correctly predicted duplexes (namely with 0 nts deviation) for both tools, estimated on all four corners of the duplexes or each individual end independently.

In order to assess the generalization performance of our tool and to compare it with *MaturePred* we use a blind, hold-out set, of 220 hairpins. Specifically, we train our model with a subset of *MaturePred*’s training set as described above (Section IV) and test its performance on a subset of hairpins

TABLE I. ACCURATE PREDICTIONS (0 NTS DEVIATION) FOR *DUPLEXSVM AND TRIVIAL* ON THE CROSS VALIDATION SET

	Accurate Predictions (%)-Cross Validation				
	k55	***k53***	***k35***	***k33***	***All Corners***
Duplex SVM	49 ± 3.7	39 ± 3.3	47 ± 4.2	41 ± 4	19 ± 1.6
Trivial	15 ± 2.7	17 ± 4.4	20 ± 5.3	18 ± 4.1	1.2 ± 1.1

from *MaturePred*'s testing set which consists of the experimentally verified animal pre-miRNAs added to miRBase versions 15-17 (4314 hairpins). In order to be able to evaluate the model's performance on every corner of the duplex, we filter out entries with unknown duplexes, multi-branch structures, and all human mouse hairpins that were used during parameter optimization in the earlier version of the algorithm. We end up with 1491 hairpins out of which we randomly select 220 hairpins for testing all models. The size of the test set was relatively small in this case solely due to time limitations. For each hairpin in the hold-out test set, we produce all candidate duplexes and select the one with the highest SVM score as the predicted candidate. We compare our model against *MaturePred* [9] and the respective *Trivial*.

MaturePred is a state-of-the-art method that employees a SVM classifier to predict the region which is most likely to contain the mature miRNA molecule in each strand of a hairpin. It has been developed using plant miRNA sequences but was also shown to perform well on several animal species [9]. We compare the tools by assessing the performance of the online *MaturePred* tool on our hold-out set, while setting the miRNA length to 22 as recommended on the web site. Since *MaturePred* predicts more than one matures (the top 10 candidates within a certain region) in each precursor's strand, while we predict only the most likely candidate, for this comparison we only consider the highest scoring candidate provided by *MaturePred*, which lies on one of the two strands. Specifically, for each candidate produced by *MaturePred*, we compare the three methods' (*DuplexSVM*, *MaturePred* and *Trivial*) performance on the strand that contains this candidate alone (ignoring the miRNA* candidate), by counting errors only on the two ends of the produced candidates.

Table II and Fig. 3 report the respective errors for each end of the predicted miRNAs that may lie on either one of the two strands, separately. For each corner *DuplexSVM,* exhibits a much higher performance than both *MaturePred* and the *Trivial* locator.

Fig. 4 shows the cumulative distribution over all hairpins of the sum of the absolute error (SAE) on all four ends of the duplex. The distributions for the three compared methods on the hold out test set are displayed. In this case, as the error is reported on all four ends, and *MaturePred* generates independent predictions for each strand, we consider the highest scoring predictions per strand as the two sides of a hypothetically predicted duplex.

TABLE II. ACCURATE PREDICTIONS (0 NTS DEVIATION) FOR *DUPLEXSVM* – *TRIVIAL* – *MATUREPRED* ON THE HOLD OUT SET

	Accurate Predictions (%) – Hold Out				
	k55	***k53***	***k35***	***k33***	***All corners***
DuplexSVM	57	24	52	35	7.2
Trivial	18	10	20	12	0.4
MaturePred	18	20	12	9	0

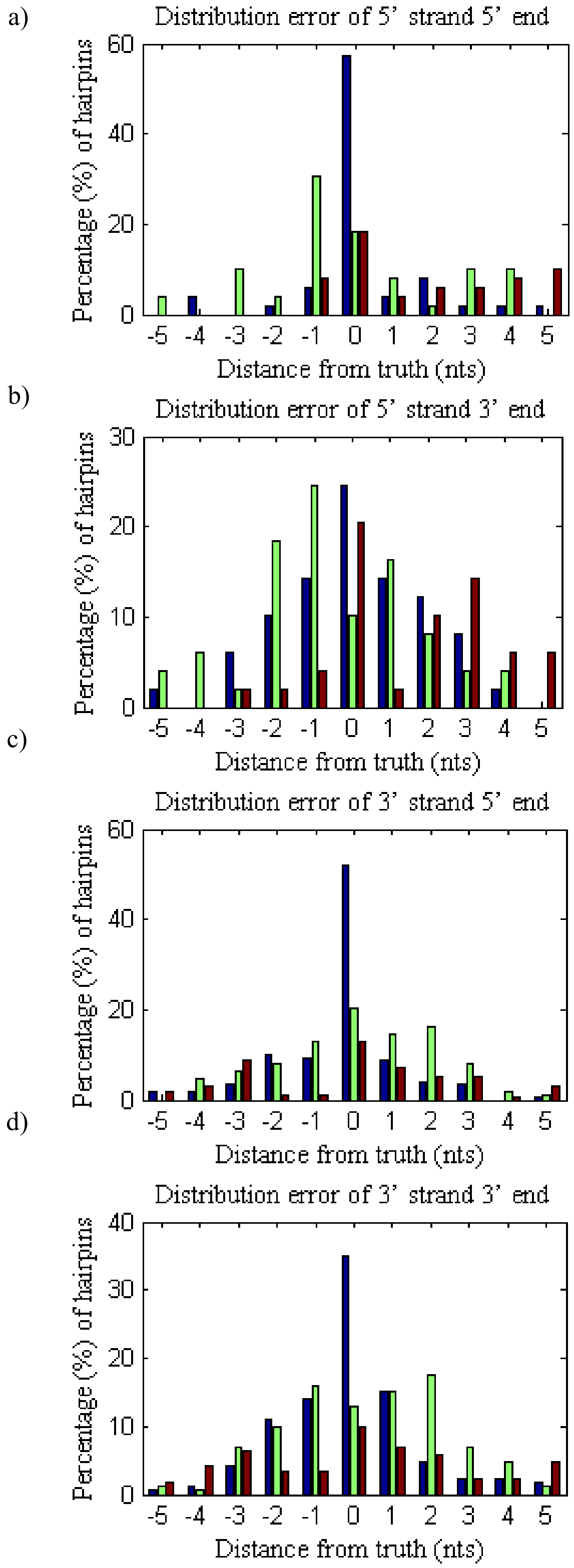

Figure 3. Distribution of prediction errors made by DuplexSVM -blue- , Trivial -green- and MaturePred -red- for each end independently: a) 5' strand 5' end, b) 5' strand 3' end, c) 3' strand 5' end, d) 3' strand 3' end. The distances between the predicted and the true positions are shown in the range of -5 to +5 nucleotides.

Importantly, Fig. 4 and Table II show that at zero nts deviation, *DuplexSVM* accurately predicts 7.2% of all duplexes while the *Trivial* and the *MaturePred* classifiers predict only 0.4% and 0% respectively. It is also worth noticing that *DuplexSVM* reaches its maximum performance much faster than *Trivial* or *MaturePred.* As is evident in Fig. 4 *DuplexSVM* reaches ~80% of accurate prediction in only 8nts deviation while *Trivial* and *MaturePred* reaches the same levels of accuracy at 11 and 35 nts (data not shown) deviation, respectively. In addition, the mean absolute error over all ends for the *DuplexSVM* is 1.61 nts with a standard deviation of 2.24 nts. This means that, on average, each end of the true duplex is identified with less than a 2 nts error. On the other hand, the mean absolute error for *MaturePred* is 5.90 nts with a standard deviation of 3.05 nts, and for the *Trivial* is 2.09 nts with a standard deviation of 2.15 nts. The observed differences are statistically significant (Wilcoxon test, p=1.0277e-008 and p=1.9917e-046 for *DuplexSVM* v.s. *Trivial* or *MaturePred*, respectively). It should be noted however that as *MaturePred* was not trained to predict duplexes, and we use its strand-independent predictions to formulate such duplexes, this comparison maybe biased in favor of our model.

Finally, Table III contains the correlation coefficients between the four end errors for the *DuplexSVM*. Note that the correlation between the error at the 5' end and the error at the 3' end is 0.81 for the 5' strand and 0.86 for the 3' strand of the duplexes. This finding indicates that the duplex and its length, are often found correctly by *DuplexSVM*, but its actual position on the hairpin can be slightly shifted. The lengths for both strands of the duplexes are predicted accurately (0 nts deviation) for 12.2% of the hairpins tested, while for each strand independently this value reaches 30.45% for 5' strand and 35.91% for 3' strand. Correlation coefficient values imply that errors in all four ends, which result from a small rightward or leftward shift, should only be counted once (and not four times as currently done), leading to a much higher performance accuracy. We are currently working on implementing such a correction.

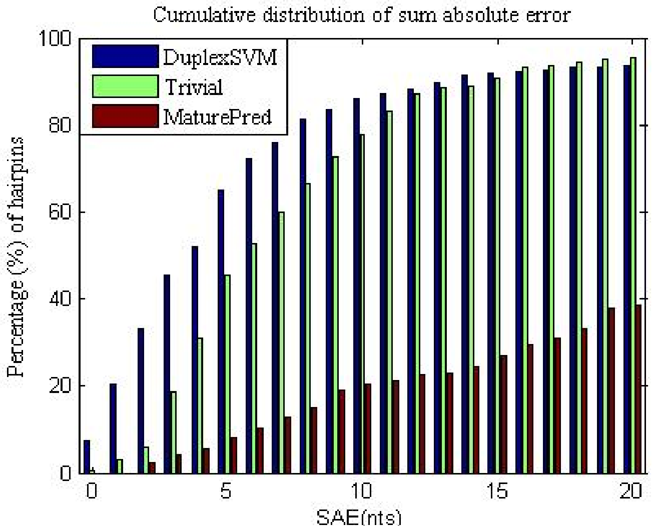

Figure 4. Cumulative distribution of the sum of the absolute error – SAE – taken over all four ends of the predicted (DuplexSVM/Trivial) or hypothetical (MaturePred) duplex; Errors up to 20 nts are displayed. The absolute difference between the true and predicted location of all of the four ends of the duplex are summed together and their cumulative distribution is being displayed.

TABLE III. CORRELATION COEFFICIENTS BETWEEN THE ERRORS IN PREDICTING K55, K53, K35, K33

	Correlation coefficients			
	Error at k55	***Error at k53***	***Error at k35***	***Error at k33***
Error at k55	1.0000	0.8101	-0.64106	-0.67465
Error at k53		1.0000	-0.66131	-0.62623
Error at k35			1.0000	0.86439
Error at k33				1.0000

VII. Discussion

In this paper we introduce the problem of predicting the miRNA:miRNA* duplex stemming from a miRNA hairpin precursor as a first step in identifying the mature miRNA(s); the latter is important both for experimentally verifying the miRNA and for computationally predicting target mRNAs. We employ biological knowledge and constraints in converting the problem to a classification one and train a high-order polynomial SVM model to identify the true duplex among candidates. The mean average error per duplex end of our predictions is 1.61 ± 2.24 nts, which highly improves against the state-of-the-art tool *MaturePred* (5.90 ± 3.05 nts); however the latter was not designed to predict miRNA duplexes. Even on a comparison biased towards *MaturePred* though, *DuplexSVM* is found to outperform the first tool by precisely identifying a larger percentage of true miRNA start positions (57% vs. 18% on the 5' strand and 52% vs. 12% on the 3' strand).

We observe large differences between the performance of *MaturePred* as reported in the original publication and our results. Specifically, we find that *MaturePred* has much lower prediction accuracy on the hold out data set used here than the original reported values. One possible explanation for this discrepancy might be the different metrics we use to compare the two algorithms. *MaturePred* [9] utilizes a "top 10" methodology, according to which the 10 highest scoring miRNA candidates for each pre-miRNA formulate their prediction. The distances between each one of the top 10 candidates and the actual miRNA are calculated and the minimum distance is reported as the prediction position deviation. This is very different from selecting only the highest scoring candidate per strand and estimating its distance from the truth, which is what we do here.

Moreover, out tool predicts both the start and the end positions of the miRNA:miRNA* sequences, whereas *MaturePred* uses a fixed length. By reporting the error on both of these corners, our findings show that length is an important feature of miRNA molecules that should be taken into account by computational tools and *DuplexSVM* is highly successful in predicting this feature.

It is worth noting that, while our tool is specifically trained to recognize miRNA:miRNA* duplexes, the comparison with

MaturePred reported here shows that learning duplexes is a very successful strategy for also identifying strand specific miRNAs. These findings suggest that considering the main biological process of miRNA biogenesis, whereby a duplex is form first and the functional strand(s) is subsequently selected, can be very useful for generating a successful computational model.

Finally, the finding that a flanking region of 14 nts from the miRNA duplex optimizes prediction performance suggests some kind of a regulatory mechanism that needs to have access to these nucleotides in order for the processing to occur. Experimental investigations are needed to explore this hypothesis further.

In summary, this is, to the best of our knowledge, the first tool capable of accurately predicting the actual miRNA:miRNA* duplex, by providing both start and end position predictions in both strands, without utilizing a fixed size window and without making the assumption of a fixed 2 nts overhang length. Future work aims to further improve the performance and generalization capacity of our algorithm by adding more features, applying our model to more species and comparing its performance to additional state-of-the-art tools.

Authors' contributions

Angelos P. Armen and Nestoras Karathanasis contributed equally to this work.

References

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," Cell, vol. 116, pp. 281-97, Jan 23 2004.

[2] A. Vermeulen, L. Behlen, A. Reynolds, A. Wolfson, W. S. Marshall, J. Karpilow, and A. Khvorova, "The contributions of dsRNA structure to Dicer specificity and efficiency," Rna, vol. 11, pp. 674-82, May 2005.

[3] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," Cell, vol. 136, pp. 215-33, Jan 23 2009.

[4] M. Yousef, L. Showe, and M. Showe, "A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification," Febs J, vol. 276, pp. 2150-6, Apr 2009.

[5] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, "Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier," Bioinformatics, vol. 22, pp. 1325-34, Jun 1 2006.

[6] Y. Sheng, P. G. Engstrom, and B. Lenhard, "Mammalian microRNA prediction through a support vector machine model of sequence and structure," PLoS One, vol. 2, p. e946, 2007.

[7] J. W. Nam, K. R. Shin, J. Han, Y. Lee, V. N. Kim, and B. T. Zhang, "Human microRNA prediction through a probabilistic co-learning model of sequence and structure," Nucleic Acids Res, vol. 33, pp. 3570-81, 2005.

[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] P. Xuan, M. Guo, Y. Huang, W. Li, and Y. Huang, "MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs," PLoS One, vol. 6, p. e27422.

[10] K. Gkirtzou, I. Tsamardinos, P. Tsakalides, and P. Poirazi, "MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors," PLoS One, vol. 5, p. e11843, 2010.

[11] Y. Wu, B. Wei, H. Liu, T. Li, and S. Rayner, "MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences," BMC Bioinformatics, vol. 12, p. 107.

[12] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," Nucleic Acids Res, vol. 39, pp. D152-7, Jan.

[13] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2.