A bioinformatics approach for investigating the determinants of Drosha processing

Nestoras Karathanasis, Ioannis Tsamardinos and Panayiota Poirazi

Abstract— **We use a bioinformatics approach to search for the biological features that determine the cleavage site of the Microprocessor complex (or Drosha) within known miRNA hairpins. Towards this goal, we employ a previously developed methodology, termed DuplexSVM, which can accurately identify the four ends of a miRNA:miRNA* duplex. Here we use DuplexSVM to study how the Drosha determines its cleavage site. We perform in silico mutagenesis experiments on 142 hairpins by changing the distance of the Drosha site from the loop tip or the stem – single stranded tails junction by adding or removing matching nucleotides. Our results suggest that the Drosha cleavage site is mainly determined by its distance from the terminal loop tip.**

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 (pri-miRNA) and 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).

Two models have been proposed on how a pri-miRNA is processed to produce a pre-miRNA. According to the first model, Drosha or the holoenzyme with Drosha providing the catalytic activity, selects an RNA hairpin bearing a terminal loop that is no less than 10 nucleotides long, and cuts ~22 nucleotides from the terminal loop – stem junction to produce a pre-miRNA [2]. According to the second model, the cleavage site is determined mainly by the distance (~11 base pairs) from the stem – single stranded tails junction [3]. 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 [4]. Irrespectively of its production process, 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 [5], 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].

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 a previous study, we introduced the problem of identifying the miRNA:miRNA* duplex as a first step in identifying the mature miRNA. We adopted 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). We showed that our tool significantly outperformed the state of the art tool MaturePred, as well as a Trivial locator, when assessed on a common blind test set[6].

Here, we use DuplexSVM to investigate the effects of mutagenesis on Drosha processing, leading to several experimentally testable predictions. In silico mutagenesis experiments performed on 142 hairpins suggest that both the distance from the single stranded stem junction and the distance from the terminal loop determine the Microprocessor's cleavage site with the latter playing a

Manuscript received November 30, 2013. This research has been co-financed by the EuropeanUnion (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.

Panayiota Poirazi is with the Institute of Molecular Biology and Biotechnology, FORTH, Greece 70013 (corresponding author, phone:00302810391139; e-mail: poirazi@imbb.forth.gr).

Nestoras Karathanasis is with the Institute of Molecular Biology and Biotechnology, FORTH, Greece 70013 and with the Department of Biology, University of Crete, Heraklion, 71409, Greece (e-mail: nk3932@hotmail.com).

Ioannis Tsamardinos is with the Department of Computer Science, University of Crete, Heraklion, 71409, Greece and with the Institute of Computer Science, FORTH, Heraklion, 70013, Greece (e-mail: tsamard@ics.forth.gr)

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

major role.

II. MUTAGENESIS PROCESS

As mentioned earlier, there are two biological models on how the microprocessor complex recognizes and process a pri-miRNA. The first model suggests that the Drosha cut site is located at ~22 nucleotides from the terminal loop – stem junction [2] while the second model claims that the cleavage site is located at ~11nts from the stem – single stranded tails junction [3]. We use our algorithm, DuplexSVM [6] to investigate concordance with these two hypotheses by performing *in silico* mutagenesis experiments. It is important to mention that, these junctions are not easy to define based on the secondary structure of miRNA hairpins. It has been shown experimentally that 8 out of 10 times, the miRNA hairpin's secondary structure, and especially their loop size and actual folding, is different from its computational prediction as shown by chemical and enzymatic probing [7]. Taking into consideration these inconsistencies and in accordance with Han et.al[3], we define two main regions on a given hairpin: region L, which includes 13 (upstream) and 11 (downstream) nucleotides from the Drosha site, and region U which includes all nucleotides between the Drosha cleavage site and the terminal loop tip, as shown in Fig. 1. Our approach relies on the tip of the loop, and not its starting position, in order to avoid errors that may have been introduced during the secondary structure generation as discussed above.

A hairpin consists of a double-stranded part, the stem, and a sequence of unmatched nucleotides that connects the strands of the stem, 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. 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 [6].

To calculate the tip we identify the last matching nucleotides before the tip, which correspond to the loop start and loop end position, respectively. If the tip is T and the last matching nucleotides are X for 5' strand and X' for 3' strand, then X < T < X' and T = X + ceil ((X'-X) / 2), ceil refers to rounding toward positive infinity.

In addition, *prediction error* is assessed using Drosha Corner Sum Absolute Error, DCSAE. The DCSAE is the sum of absolute errors in number of nucleotides from true position between the actual and the predicted Drosha site end, taken over both ends of the Drosha site. For example, if the true positions of Drosha site are k55 = XX55 and k33 = XX33 and the predicted positions are YY55, and YY33 respectively, then the DCSAE = $|XX_{55} - YY_{55}| + |XX_{33} - YY_{33}|$.

In order to characterize the effect of nucleotide mutations on Drosha processing we used all human and mouse hairpins in miRBase 19.0 as graphically illustrated in Fig. 1. To ensure the presence of the stem – single stranded tails junction in our sequences and the sequence - structure around it, we added whenever needed, 23 (upstream) and 21 (downstream) nucleotides from the Drosha site [3]. Out of this dataset, DuplexSVM was trained with the same hairpins that were used during parameter optimization [6] (678) and the remaining 383 hairpins were used for testing.

Only 142/383 hairpins whose Drosha sites were predicted correctly (0nts deviation) were used for the mutagenesis experiments.

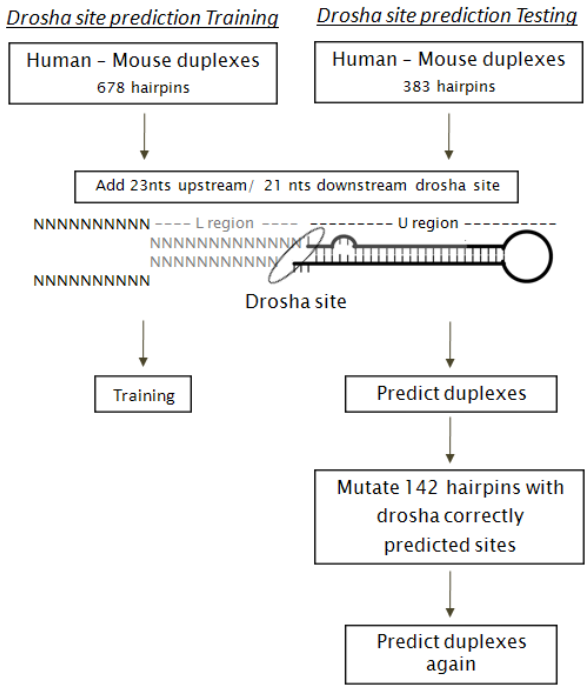

Fig. 1. Flowchart of the in silico mutagenesis process. The L region corresponds to 13nts before and 11 nts after the Drosha cut site. The U region corresponds to the hairpin starting at the Drosha site and containing the loop.

In silico mutagenesis was performed by inserting or deleting 2, 4 or 6 matching nucleotides to L or U regions, thus shifting the Drosha site towards or away from the single stranded tails – stem junction, or the terminal loop tip. DuplexSVM was re-applied on the mutated hairpins and the new Drosha processing sites were predicted, Fig. 1.

III. RESULTS

In order to evaluate if a mutation has a statistical significant effect on the DuplexSVM's predictions we perform Wilcoxon rank-sum tests between the results obtain using the Drosha Corners Sum Absolute Error (DCSAE) before and after every mutation. Bonferroni correction was also applied to correct for multiple testing.

Table one shows the statistically significant results from this comparison. The Drosha Site Average Shift (DSAS) and the percentage of hairpins in which a shift was observed after each mutation are also shown. In order to calculate the DSAS, only hairpins whose Drosha site had changed after a given mutation were taken into account. For each hairpin we first calculate Drosha Corners Mean Error, DCME = $((YY_{55} - XX_{55}) + (XX_{33} - YY_{33}))/2$. Subsequently we estimate the median value over all hairpins' DCMEs, DSAS = median(DCMEs). As evident from the table, all mutations in the U region (between the Drosha site and the loop tip) resulted in a predicted shift in the Drosha site while for the L region (between the Drosha site and the stem – single stranded tails junction), only the deletion of 4nts led to significant changes.

Specifically, the deletion of 4 nucleotides in the L region resulted in a 0.5nt shift of the Drosha site towards the stem – single stranded tail junction (see Table 1). This finding is in contrast with recent experimental work whereby the deletion

TABLE I
MUTATION ANALYSIS

Type of mutations	DSAS	Percentage %	Significance
L-4	-0,5	9.75	***
U+2	0.5	15.44	***
U+4	2.5	19.51	***
U+6	5	20.32	***
U–2	-1.5	20.32	***
U–4	- 4.5	43.9	***
U–6	-6.5	78.86	***

In silico mutagenesis experiments. The first column lists the type of the performed mutations, i.e. the number of matching nts added (+) or deleted (-) in the L or U regions. The sequence of the inserted nucleotides was generated randomly. The second column shows the Drosha Site Average Shift (DSAS) in nts which corresponds to the median of the Drosha Corners Mean Error. The third column reports the percentage of hairpins in which a shift was predicted. The last column shows the statistical significance of these effects, assessed using a. Wilcoxon rank-sum test between DCSAE calculated on wild type and mutated sequences for each mutation. *** corresponds to p-value ≤ 0.001

of 4 matching nucleotides in the respective L region of mir-16-1, pushed the Drosha site away from the stem – single stranded tail junction by the same distance[3]. A thorough analysis of our results, revealed that for an approximate 60% of the cases the Drosha cleavage site moves closer to the stem – single stranded tail junction by 1 nucleotide and the remaining 40% of the times it moves away by 2 nucleotides (see Table 2). Deletion of nucleotides form the U region however results in a shift of the Drosha cleavage site which is analogous to the direction of the mutation: inserting or deleting nucleotides moves the Drosha site in a way that maintains specific distance from the stem loop tip. These findings are in close agreement with the experimental work of Yan Zeng et al [2].

IV. CONCLUSION

Using a state-of-the-art computational method for the identification of miRNA:miRNA* duplexes [6], this study explored the effect of mutations on determining the Drosha cleavage site. Our findings are in agreement with certain experimental data, suggesting that such a bioinformatics approach can be used to investigate the rules underlying Drosha processing.

TABLE II
L REGION MUTATIONS

Type of mutations	DSAS	Percentage %
L-4	-1	58.3
L-4	2	41.6

The effect of L-4 type of mutations is shown. In 58% of the cases presenting an effect from this mutation the Drosha site moved one nucleotide from its original place towards the single stranded stem junction. In the remaining 42% of the cases, it moved 2 nucleotides towards to the stem loop junction.

Specifically, there are currently two biological models regarding the determination of the Drosha cleavage site: according to the model of Han et al [3], the complex cuts ~11nts from the stem – single stranded tails junction. According to the model of Zeng et al [2, 8], the microprocessor complex recognizes and cleaves a pri-miRNA ~22nts from the stem – loop junction. Our results suggest a third model that combines information from both of these studies.

In silico addition and/or deletion of matching nucleotides showed that the region after (U region in Fig. 1) is more important than the region before (L region in Fig. 1) the Drosha processing site. Specifically, every mutation in the U region resulted in a shift of the Drosha cleavage site while only the deletion of 4 in the L region had a similar effect. Our findings are in agreement with the recent work of Vincent et al. [9], where the Microprocessor complex was shown to distinguish between hairpins from different species by relying on sequence motifs that lie either within the single stranded tail region or the loop region (U region). These motifs have been suggested to guide the Microprocessor complex towards its cleavage site. It is possible that the insertion/deletion of nucleotides in the L and U regions that we simulate with DuplexSVM alters either the motifs themselves or the distances between the motifs and the Drosha cleavage site, thus resulting in the observed shift in the cleavage site itself. In sum, both ours and previous experimental findings suggest that structural and sequence information on both sides influence the Drosha cleavage point.

REFERENCES

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

[2] Y. Zeng, *et al.*, "Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha," *EMBO J,* vol. 24, pp. 138-48, Jan 12 2005.

[3] J. Han, *et al.*, "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex," *Cell,* vol. 125, pp. 887-901, Jun 2 2006.

[4] V. N. Kim, *et al.*, "Biogenesis of small RNAs in animals," *Nat Rev Mol Cell Biol,* vol. 10, pp. 126-39, Feb 2009.
[5] A. Vermeulen, *et al.*, "The contributions of dsRNA structure to Dicer specificity and efficiency," *Rna,* vol. 11, pp. 674-82, May 2005.
[6] N. Karathanasis, "SVM-based miRNA: MiRNA∗ duplex prediction," 2012, pp. 181-186.
[7] J. Krol, *et al.*, "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, Oct 1 2004.
[8] X. Zhang and Y. Zeng, "The terminal loop region controls microRNA processing by Drosha and Dicer," *Nucleic Acids Res,* vol. 38, pp. 7689-97, Nov 2010.
[9] V. C. Auyeung, *et al.*, "Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing," *Cell,* vol. 152, pp. 844-58, Feb 14 2013.