Finding Motifs in miRNA Sequences

Finding Motifs in miRNA Sequences

Chinnu Subramaniam [1], Jiyuan An [1], Jayavardhana Gubbi [2] and Yi-Ping Phoebe Chen [1]
[1] *School of Engineering and Information Technology, Deakin University, Australia*
[2] *Department of Electrical and Electronic Engineering, The University of Melbourne*
{subbu, jiyuan, phoebe }@deakin.edu.au
jrgl(@)ee.unimelb.edu.au

Abstract

Recently, the emergence of non-coding miRNA has attracted biology and computer researchers. miRNA plays an important role in regulation of genes. Finding motifs in RNA is one of important topics. In our work, we attempt to find motifs in mature miRNA from combinations ranging from two to ten nucleotides. Interestingly, we have found several motifs only appear in mature miRNA but not appear in other regions of primary miRNA sequences taken from latest miRNA datasets. The findings of our investigation may help in the building model to predict all possible miRNAs in genomes

1. Introduction

MicroRNAs (miRNAs) are non-coding RNA genes and have 21 or 22 nucleotides in the sequences [1]. They are found to regulate gene expressions in several diseases such as cancer [2, 3] and diabetes [4]. Regulation or translation is repressed by binding to complementary regions of messenger transcripts [5]. miRNAs participate in diverse roles, namely, 1) developmental timing, 2) cell death and fat metabolism, 3) haematopoiesis in mammals, and 5) leaf development and floral patterning in plants [6].

The mechanism of miRNA transcription is fairly well understood and can be explained as shown in Fig. 1. Polymerase II is used to transcribe the primary miRNA (pri-miRNAs) [4]. The transcribed pri-miRNA is further processed using the enzyme called Drosha endonuclease in the nucleus [5]; it is then exported to the cytoplasm with the help of the enzyme, Exportin 5 in the form of pre-miRNA stem loops [6]. In the cytoplasm, the miRNA is produced using Dicer endonuclease [7]. A stem loop secondary structure is found to be crucial in this mechanism [8].

Identification of miRNA is crucial in understanding the intricacies of the interaction of different parts of gene such as intron, exon and intergenic regions. Intensive effort has been focussed on prediction of miRNA using computation methods.

Finding motifs or patterns from miRNAs and pre-miRNAs is important in molecular cell biology as it paves way for identifying all possible miRNA targets of genomes so that the mechanism that controls the pathway and the gene expression can be well understood. It is projected that the maximum of miRNA could be around 1000 in mammalian genomes [7]. However, experimental verification needs rigorous effort from wet labs.

In this paper, we have generated all possible motifs and ranked the identified motifs based on an index defined in this work. In section 2, the algorithm used in this work is discussed. The section 3 explains the data used and methods of analysis. In section 4, results are presented.

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

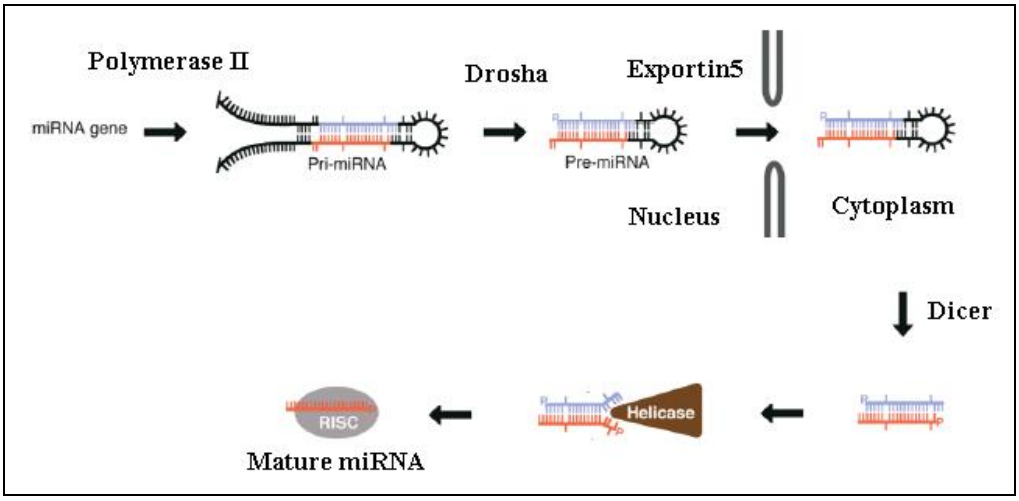

Fig.1.TranscriptionofmiRNA

2. Algorithm

We use a robust force algorithm to calculate the number of primary miRNA and mature sequences that include a pattern. If a pattern appears in a sequence, we can say that the sequence supports the pattern. Throughout the paper, we will use support to represent this number. For example, a pattern appears in two sequences of the dataset, we can say that the pattern has two supports. To find motif of mature miRNA sequence, we want to find nucleotide combinations that only appear in mature miRNA sequence, but not appear in primary miRNA.

A naïve way to find mature miRNA motif is enumerating all possible combinations of 4 nucleotides (A,G,C,U) and calculate their supports. However, the number of the combinations increases exponentially. The computation is costly in both CPU time and memory space. To deal with thousands of miRNA, an effective method is needed.

We notice the fact that the number of supports will not increase when a combination is expanded. For example, if combination "ACCU" has 135 supports, the combination "ACCUA" has equal to or less than 135 supports. Therefore, for a given threshold number of supports for a mature miRNA, if a combination has less than threshold of supports, its all expanded combinations has less than threshold of supports as well. This kind of combinations can be pruned away in support computation. Note that it will save a large amount of CPU time and memory space for support computation. The program code is shown in the figure 2 below.

```
1.   For 1st = {A, G, C, U}
2.       num1(1st) = support(1st)
3.       if num1(1st) < threshold
4.          continue for
5.       end if
6.       For 2nd = {A, G, C, U}
7.          num2(1st+2nd) = support(1st+2nd)
8.          if num2(1st + 2nd) < threshold
9.             continue for
10.         end if
11.         …..
12.            for nth = {A, G, C, U}
13.                numn(1st+2nd+..nth)=support(1st+2nd+…+nth)
14.            next
15.    next
16.  next
```

Fig2.Algorithm

3. Materials used in the Experiment

The hairpin and mature sequences were downloaded from the release 7.1 of the MicroRNA repository, Rfam7.1. All the 3424 sequences of the available species are used in our analysis [8].

4. Result and Discussion

The motifs have been identified using brute force algorithm as shown in Fig. 1. It is found that two types of motifs occur in the primary sequences, namely, 1) motifs that occur in the same type of genes with variation homologically and 2) motifs that occur in different genes as shown in Appendix.

Further, a motif occurrence ratio (MOR) is used to identify the importance of the motif in locating miRNA position. It is calculated by dividing the number of occurrence of the motif with that of the hairpin sequence. If the value is one, it indicates that the motif occurs only at the mature RNA region. It can be used to eliminate the false positive when a supervised learning is made. On the other hand, if the value is less than one, it indicates that the motifs occur also at different locations of the hairpin sequence. Furthermore, if the value is very low, it provides information on its being excluded in mature RNA. It is also interesting one as it can be used to eliminate the regions that do not have the potential miRNA.

Table 1. Motifs and their occurrences in mature (M) and hairpin (H) sequences

10 Nucleotides	M	H	MOR	9 Nucleotides	M	H	MOR	8 Nucleotides	M	H	MOR
AUGACUUGCC	70	70	1.0000	AGCCAAGGA	79	80	0.9875	AAAGUGCU	92	139	0.6619
AGCCAAGGAU	68	68	1.0000	AUGACUUGC	71	71	1.0000	GACUUGCC	82	88	0.9318
CCAAGGAUGA	68	68	1.0000	UGAGGUAGU	70	97	0.7216	GCCAAGGA	79	81	0.9753
GCCAAGGAUG	68	68	1.0000	UGACUUGCC	70	71	0.9859	AGCCAAGG	79	80	0.9875
GGAUGACUUG	67	69	0.9710	CUUCAUUCC	69	72	0.9583	UGAGGUAG	77	108	0.7130
AAGGAUGACU	67	67	1.0000	CAAGGAUGA	68	68	1.0000	GAGGUAGU	74	102	0.7255
AGGAUGACUU	67	67	1.0000	CCAAGGAUG	68	68	1.0000	UUCAUUCC	72	75	0.9600
CAAGGAUGAC	67	67	1.0000	GCCAAGGAU	68	68	1.0000	AUGACUUG	72	74	0.9730
GAUGACUUGC	67	67	1.0000	AGGAUGACU	67	70	0.9571	UGACUUGC	71	72	0.9861
UGAGGUAGUA	62	86	0.7209	GAUGACUUG	67	69	0.9710	AAGUGCUU	69	113	0.6106
CGGACCAGGC	61	61	1.0000	GGAUGACUU	67	69	0.9710	GUGCAGGU	69	86	0.8023
GGACCAGGCU	61	61	1.0000	AAGGAUGAC	67	68	0.9853	CUUCAUUC	69	79	0.8734
UCGGACCAGG	61	61	1.0000	GAGGUAGUA	65	89	0.7303	AAGGAUGA	68	72	0.9444
CCAGGCUUCA	60	61	0.9836	GUGCAGGUA	64	68	0.9412	CAAGGAUG	68	70	0.9714
ACCAGGCUUC	60	60	1.0000	CAGGCUUCA	62	63	0.9841	CCAAGGAU	68	68	1.0000
GACCAGGCUU	60	60	1.0000	CGGACCAGG	61	61	1.0000	GGAUGACU	67	73	0.9178
AAGCUGCCAG	59	60	0.9833	GACCAGGCU	61	61	1.0000	AGGAUGAC	67	71	0.9437
UGACAGAAGA	57	57	1.0000	GGACCAGGC	61	61	1.0000	GAUGACUU	67	70	0.9571
CAGGCUUCAU	56	56	1.0000	UCGGACCAG	61	61	1.0000	AGGUAGUA	65	92	0.7065
AGGCUUCAUU	54	54	1.0000	AAGCUGCCA	60	61	0.9836	UGCAGGUA	64	70	0.9143
GCUUCAUUCC	54	54	1.0000	CCAGGCUUC	60	61	0.9836	AGGCUUCA	62	65	0.9538
GGCUUCAUUC	54	54	1.0000	ACCAGGCUU	60	60	1.0000	CAGGCUUC	62	63	0.9841

GAGGUAGUAG	52	76	0.6842	UAAACAUCC	59	69	0.8551	UCGGACCA	61	64	0.9531
AGUGCAGGUA	51	54	0.9444	AGCUGCCAG	59	64	0.9219	UGACAGAA	61	63	0.9683
GUAAACAUCC	50	60	0.8333	GACAGAAGA	59	59	1.0000	CGGACCAG	61	62	0.9839
UGUAAACAUC	50	60	0.8333	AAAGUGCUU	58	68	0.8529	GGACCAGG	61	62	0.9839
UAGCCAAGGA	49	49	1.0000	UGACAGAAG	57	58	0.9828	ACCAGGCU	61	61	1.0000
GUGCAGGUAG	47	67	0.7015	AGGCUUCAU	56	56	1.0000	GACCAGGC	61	61	1.0000
ACAGAAGAGA	47	47	1.0000	GCUUCAUUC	54	54	1.0000	AGCUGCCA	60	72	0.8333
CAGAAGAGAG	47	47	1.0000	GGCUUCAUU	54	54	1.0000	AAGCUGCC	60	61	0.9836

5. Conclusion

In this paper, we propose a brute force algorithm to find motifs in mature miRNA within primary miRNA sequences. From the latest release 7.1 of the MicroRNA dataset, we have found very interesting motifs that only appear in mature miRNA. It may be used to predict all possible miRNA in genomes. In our future work, we will find out alignment of different lengths of motifs. Furthermore, the positions of occurrences show interesting features that can be used in building supervised learning model in identifying miRNA targets in genomes.

6. References

[1] Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, 116, 281–297.

[2] Wiemer, E.A.C, miRNAs and cancer, (2006) Journal of RNAi and Gene Silencing, 2(2), 173-174.

[3] He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J. and Hammond, S.M. (2005) A microRNA polycistron as a potential human oncogene, Nature, 435, 828–833.

[4] Gauthier, B.R. and Wollheim, C.B. (2006), MicroRNAs: 'ribo-regulators' of glucose homeostasis, doi:10.1038/nm0106-36, Nature Medicine 12, 36-38.

[5] Sontheimer, E.J. and Carthew, R.W. (2005) Silence from within: endogenous siRNAs and miRNAs, Cell, 122, 9–12.

[6] Griffiths-Jones, Sam, Grocock, R.J., Dongen, S.V, Bateman, A and Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature, Nucleic Acids, Research, 2006, Vol.34, Database issue, D140–D144.

[7] Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R., Cuppen, E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes, Cell, 120:21-24.

[8] Griffiths-Jones S: The microRNA Registry. Nucleic Acids Res 2004, 1(32): D109-D111.

Appendix

motif	length	# occurrence		motif	length	# occurrence
AUGACUUGCC	10	70		CCAAGGAUGA	10	68
name	**species**			**name**	**species**	
>ath-miR169a	Arabidopsis	thaliana		>ath-miR169a	Arabidopsis	thaliana
>ath-miR169b	Rattus	norvegicus		>ath-miR169b	Rattus	norvegicus
>ath-miR169c	Rattus	norvegicus		>ath-miR169c	Rattus	norvegicus
>ath-miR169d	Rattus	norvegicus		>ath-miR169d	Rattus	norvegicus
>ath-miR169e	Rattus	norvegicus		>ath-miR169e	Rattus	norvegicus
>ath-miR169f	Rattus	norvegicus		>ath-miR169f	Rattus	norvegicus
>ath-miR169g	Rattus	norvegicus		>ath-miR169g	Rattus	norvegicus
>ath-miR169h	Rattus	norvegicus		>ath-miR169h	Rattus	norvegicus

>ath-miR169i	Rattus	norvegicus	>ath-miR169i	Rattus	norvegicus
>ath-miR169j	Rattus	norvegicus	>ath-miR169j	Rattus	norvegicus
>ath-miR169k	Rattus	norvegicus	>ath-miR169k	Rattus	norvegicus
>ath-miR169l	Rattus	norvegicus	>ath-miR169l	Rattus	norvegicus
>ath-miR169m	Rattus	norvegicus	>ath-miR169m	Rattus	norvegicus
>ath-miR169n	Rattus	norvegicus	>ath-miR169n	Rattus	norvegicus
>gma-miR169	Mus	musculus	>gma-miR169	Mus	musculus
>mtr-miR169a	Human	cytomegalovirus	>mtr-miR169a	Human	cytomegalovirus
>mtr-miR169b	Human	cytomegalovirus	>mtr-miR169b	Human	cytomegalovirus
>osa-miR169a	Oryza	sativa	>osa-miR169a	Oryza	sativa
>osa-miR169b	Arabidopsis	thaliana	>osa-miR169b	Arabidopsis	thaliana
>osa-miR169c	Arabidopsis	thaliana	>osa-miR169c	Arabidopsis	thaliana
>osa-miR169e	Oryza	sativa	>osa-miR169d	Oryza	sativa
>osa-miR169f	Oryza	sativa	>osa-miR169e	Oryza	sativa
>osa-miR169g	Oryza	sativa	>osa-miR169f	Oryza	sativa
>osa-miR169h	Oryza	sativa	>osa-miR169g	Oryza	sativa
>osa-miR169i	Oryza	sativa	>osa-miR169h	Oryza	sativa
>osa-miR169j	Oryza	sativa	>osa-miR169i	Oryza	sativa
>osa-miR169k	Oryza	sativa	>osa-miR169j	Oryza	sativa
>osa-miR169l	Oryza	sativa	>osa-miR169k	Oryza	sativa
>osa-miR169m	Oryza	sativa	>osa-miR169l	Oryza	sativa
>osa-miR169n	Oryza	sativa	>osa-miR169m	Oryza	sativa
>osa-miR169o	Oryza	sativa	>ptc-miR169a	Danio	rerio
>ptc-miR169a	Danio	rerio	>ptc-miR169b	Danio	rerio
>ptc-miR169b	Danio	rerio	>ptc-miR169c	Danio	rerio
>ptc-miR169c	Danio	rerio	>ptc-miR169d	Danio	rerio
>ptc-miR169d	Danio	rerio	>ptc-miR169e	Danio	rerio
>ptc-miR169e	Danio	rerio	>ptc-miR169f	Danio	rerio
>ptc-miR169f	Danio	rerio	>ptc-miR169g	Danio	rerio
>ptc-miR169g	Danio	rerio	>ptc-miR169h	Danio	rerio
>ptc-miR169h	Danio	rerio	>ptc-miR169i	Danio	rerio
>ptc-miR169i	Danio	rerio	>ptc-miR169j	Danio	rerio
>ptc-miR169j	Danio	rerio	>ptc-miR169k	Danio	rerio
>ptc-miR169k	Danio	rerio	>ptc-miR169l	Danio	rerio
>ptc-miR169l	Danio	rerio	>ptc-miR169m	Danio	rerio
>ptc-miR169m	Danio	rerio	>ptc-miR169n	Danio	rerio
>ptc-miR169n	Danio	rerio	>ptc-miR169o	Danio	rerio
>ptc-miR169o	Danio	rerio	>ptc-miR169p	Danio	rerio
>ptc-miR169p	Danio	rerio	>ptc-miR169r	Danio	rerio
>ptc-miR169r	Danio	rerio	>ptc-miR169s	Danio	rerio
>ptc-miR169s	Danio	rerio	>ptc-miR169v	Danio	rerio
>ptc-miR169t	Danio	rerio	>ptc-miR169w	Danio	rerio
>ptc-miR169v	Danio	rerio	>ptc-miR169x	Danio	rerio
>ptc-miR169w	Danio	rerio	>sbi-miR169a	Zea	mays
>sbi-miR169a	Zea	mays	>sbi-miR169b	Zea	mays
>sbi-miR169b	Zea	mays	>sbi-miR169c	Sorghum	bicolor
>sbi-miR169c	Sorghum	bicolor	>sbi-miR169d	Sorghum	bicolor

>sbi-miR169d	Sorghum	bicolor		>sbi-miR169e	Sorghum	bicolor
>sbi-miR169e	Sorghum	bicolor		>sbi-miR169f	Sorghum	bicolor
>sbi-miR169f	Sorghum	bicolor		>sbi-miR169g	Sorghum	bicolor
>sbi-miR169g	Sorghum	bicolor		>sbi-miR169h	Sorghum	bicolor
>sbi-miR169h	Sorghum	bicolor		>zma-miR169a	Caenorhabditis	briggsae
>sbi-miR169i	Sorghum	bicolor		>zma-miR169b	Caenorhabditis	briggsae
>zma-miR169a	Caenorhabditis	briggsae		>zma-miR169c	Glycine	max
>zma-miR169b	Caenorhabditis	briggsae		>zma-miR169f	Glycine	max
>zma-miR169c	Glycine	max		>zma-miR169g	Glycine	max
>zma-miR169f	Glycine	max		>zma-miR169h	Glycine	max
>zma-miR169g	Glycine	max		>zma-miR169i	Glycine	max
>zma-miR169h	Glycine	max		>zma-miR169j	Glycine	max
>zma-miR169i	Glycine	max		>zma-miR169k	Glycine	max
>zma-miR169j	Glycine	max				
>zma-miR169k	Glycine	max				