Computational Prediction of MicroRNAs from Chromosome no. 3 & 4 of Plasmodium vivax

Ranojit Sarker1
1Dept. of biotechnology, Neotia Institute of Technology, Management and Sciences. DH Road, 24 PGS(s), India.

Ujjwal Maulik2
2Dept. of Computer Science and Engineering, Jadavpur University, kolkata, India,

Abstract— MicroRNAs (MiRNA's) are ~22nt long; one of the types of non-coding RNA has an important role in the gene regulation network, either by termination of the translation procedure or sometimes activating the gene, actively or passively. Host-Parasite relationship largely depends on the gene regulatory network as the parasite has to survive in the host cell against the immune response of the host. So miRNA may have an important role in the host parasite relationship. The malarial parasite needs to survive in the human host cell during its life cycle. As it is well known to us today that the gene regulatory network is largely controlled by the non coding miRNA, here we are interested to find out potential miRNA in the parasite genome computationally. We have identified few putative miRNA genes in the chromosome 3 & 4 of Plasmodium vivax which need to be validated by experimental procedures.

Keywords-component; miRNA prediction, non coding RNA, Host- Parasite relationship.

I. INTRODUCTION

MicroRNAs (miRNA's) are ~22nt long; one of the types of non-coding RNA having an important role in the control of the gene regulation, more specifically termination of the translation procedure. But there are also evidences of activation of gene expression by miRNA's [1]. MiRNAs are encoded by genes. They are first transcribed as long pri-miRNAs and processed to 70 to ~110 nt precursors (pre-miRNA) with stem–loop structure by the RNase III enzyme Drosha. Then, another RNase III enzyme Dicer processed the pre-miRNAs to release the 22 nt mature miRNAs. Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules. MiRNA's were first described in 1993 by Lee and colleagues in the Victor Ambros lab, yet the term miRNA was only introduced in 2001 in a set of three articles in Science. [2] MiRNAs play an important regulatory role in many cellular and developmental processes like cell division [9], cell death [10], hormone secretion [11] neural development [12] tumor suppression [13], oncogenesis [14] etc. Since the discovery of the first miRNA lin-4 [15] and let-7 [16] in Caenorhabditis elegans, presence of miRNA's in the genomes of various organisms [17,18] including insects [19], plants [20], higher vertebrates [21] and viruses [22] has been reported. MiRNAs are generally conserved in closely related species and to some extent in distantly related species as well [23]; e.g. about 10% of miRNAs identified in invertebrates are also conserved in mammals and other higher animals, suggesting cross-species conservation of their regulatory functions [24]. In case of parasites under Plasmodium genera, no data so far has been reported in the miRNA registry miRBase. In Plasmodium genera, currently over 200 species has been recognized and new species continue to be described. Of the 200+ known species of Plasmodium, at least 10 species infect humans. Among them four parasites are responsible for human malaria: Plasmodium falciparum, Plasmodium vivax, Plasmodium malaria, Plasmodium ovale. Plasmodium falciparum is the most deadly in the African region and Plasmodium vivax is mainly responsible for the malarial cases outside Africa, like India. This is one of the reasons to select P. vivax for our study.

II. METHODS

To design an algorithm for parasitic miRNA prediction may not give a very specific result, as there is no experimentally verified miRNA data from parasites. For this reason here we have adopted a procedural approach which has been classified in three categories. The part one is the pre-microRNA prediction from the chromosomes, the part two contains the filtering of collected sequences and the part three belongs to the similarity search with the miRNA database (miRBase).

A. Pre- microRNA prediction

As we stated before that there is no experimentally verified data about plasmodium (parasite), it may be erroneous approach to formulate algorithmic features considering the data from different organism. So other than defining an algorithmic feature set, we considered five available algorithms for miRNA prediction from different species. The objective is that collectively these algorithms contain number of features which will screen a number of miRNAs (including false positives), and then these putative miRNAs can be filtered using various miRNAs features.

Firstly, we considered two kinds of approaches, 'the two class approach' and 'the one class approach' to identify the potential pre-miRNA prediction. In two class approach all the algorithms are trained with both positive and negative datasets. In one class approach all the algorithms are trained with only positive class of known miRNAs.

Phase 1: (using of three different two class algorithms)

MirEval: MirEval [4] offers two separate structural analysis algorithms that use two different, non-redundant approaches. Input sequences are analyzed with a sliding window of 80 nt. Each window is evaluated for stable secondary structures by RNA-fold [25] and each hairpin shape (a helix of at least 15 nt with no internal hairpin) is then analyzed as follows. The first algorithm, 'Triplet-SVM classifier' [5], is based on support vector machines (SVM) and takes into account primary sequence and structural features to classify candidates. The features focus on the information of every 3 adjacent nucleotides, and named as triplet structure sequence features or triplet features. In the predicted secondary structure, there are only two statuses for each nucleotide, paired or unpaired, indicated by brackets ("("or")") and dots ("."), respectively. The left bracket "(" means that the paired nucleotide is located near the 5'-end and can be paired with another nucleotide at the 3'-end, which is indicated by a right bracket")". As an example, Figure 1 illustrates how a hairpin is represented using triplet features. It's exclude the terminal loop and external single stranded regions of the hairpin and only considers the stem portions. The number of appearance of each triplet element is counted for each hairpin (pre-miRNA or pseudo pre-miRNA) to produce the 32-dimensional feature vector. It is normalized before being used as input features for SVM. It is able to distinguish pre-miRNAs from other hairpin shapes with 90% accuracy.

As conserved stem–loops structures of miRNAs and other ncRNAs can be distinguished by their secondary structural features, this algorithm considered commonly used factors such as free energy and the number of consecutive base-pairs as well as more specific information on the position and size of bulges and loops. The table 1 includes the values of different parameters which used as the cutoffs for miRNA prediction.

TABLE 1 Different parameters which used as the cutoffs for miRNA prediction.

Parameter	Mean value, precursors	Mean value, nonprecursors
Bulge dist from apical loop (nt)	8.8	9.8
Internal loop dist from start (nt)	9.43	10.43
Internal loop dist from apical loop (nt)	8.74	7.80
Total number of bp	31	29.2
Bulge dist from start (nt)	16	11.45
ΔG (kcal)	-33	-29
Max consecutive bps in stem	9.38	8.33
Number of internal loops	3.12	3.89
Apical loop size (nt)	7.1	13.1
Internal loop size (nt)	1.94	3.20
Number of bulges	1.86	3.29
Bulge size (nt)	1.47	4.37

The MirEval takes up to 10000 nt genome sequence as input. It provides a species selection option in which we have selected the "other species" option. The out put of this tool is shown in fig: 1.

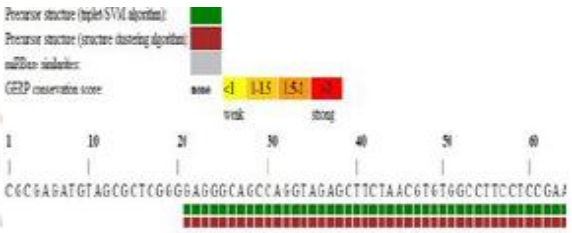

Figure1. MirEval output

In output page the predicted pre-miRNA regions are blocked by different colours like, green colour indicates the region of the genome selected by the triplet SVM algorithm and dark red colour indicates selection of structure clustering algorithm [fig: 1]. We have scanned the entire two chromosomes of Plasmodium vivax by this tool and created a table containing all the potential pre-miRNAs with specific chromosome positions. B. ProMiR 2: Like MirEval we have used another very well known probabilistic statistical tool named ProMiR 2 [7], which is an upgraded version of ProMiR. ProMiR has been used successfully to predict a miRNA in a stem–loop sequence using a score generated by a probabilistic co-learning model. But this improved method was developed to identify the conserved and non-conserved miRNAs near known miRNAs or candidates. This strategy is very useful because more than half of the known miRNA genes are present as tandem arrays within operon-like clusters. This new version, ProMiR II, generates a list of nearby potential miRNAs according to score and to several filtering criteria such as conservation score, entropy, G/C ratio and free energy. This enhanced method allows for low- or high stringency prediction of conserved and non-conserved miRNA genes by adjusting the filtering criteria. ProMiR 2 is a completely statistical tool which provide three different programs, (a) ProMiR-v; search for potential miRNAs in the Vicinity of known miRNAs, (b) ProMiR-c; search for potential miRNAs in the vicinity of a Candidate, (c) ProMiR-g; predict miRNAs in a long sequence, a Generalized version of ProMiR. We have used ProMiR-g for pre- miRNA prediction. It takes 10000 nt genome sequence as a input. It does not have any option for other species. So we have chosen C.elegans from species selection option because C.elegans is taxonomically closer to the Plasmodium species than others. This tool uses the following filtering cutoffs for pre-miRNAs prediction: (a) Input parameters: window size: 100 nt; shift size: 10; ProMiR value: 0.033. (b) Filtering parameters: conservation score: >- 0.0; free energy: >=-25 kcal/mol; G/C ratio: 0.3~0.7; entropy: >= 1.8. Like results of MirEval we have created a table for ProMiR

2-g, which contains the predicted sequences with the specific positions.

Phase 2: (using three different one class algorithms) in one class approach we have used "The One Class miRNA classifier program" [8]. It has five different one class classifier algorithms from which we have used following three algorithms. (a) OC-SVM, (b) OC-Gaussian, (c) OCKNN. All these three algorithms are trained by only the positive data sets of known miRNAs. After phase 1 we have created a combined table with sequences from all the two class algorithms. Then we have verified every pre-miRNA sequence by above mentioned one class algorithms. This classifier describes features of miRNAs extracted from both secondary structure and sequences. For the positive (miRNA) class, the 21 nt of the mature miRNAs are mapped into its associated stem-loop (generated by the mfold program) and then features are extracted as described below. For the structural features, 62 features are derived from three parts of the associated hairpin (stem-loop) (fig.4) – foot, mature, and head – and include the following for each of these parts: (1) the total number of base pairs (bp), (2) the number of bulges, (3) the number of loops, (4) the number of asymmetric loops, (5) eight features representing the number of bulges of lengths 1–7 and greater than 7, (6) eight

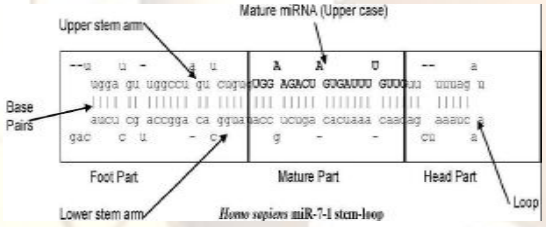

Figure 2. Partition stem-loop into 3 parts, foot, mature and head features to determine potential stem-loops.

features represents the number of symmetric loops of length 1–7 and greater than 7, (7) the distance from the mature miRNA candidate to the first paired base of the foot and head part. For the sequence features, the classifier defines "words" as sequences having lengths equal to or less than 3. The frequency of each word in the first 9 nt of the 21 nt putative mature miRNA is extracted to form a representation in the vector space. This system of "words" has been done with all the one class algorithms

Now, these three results of a single sequence describe the different approach of scoring the sequences. Otherwise, the selected portion represented as precursor is same. Nevertheless, the miRNA results are different. Therefore, it requires more detail study to find mature miRNAs.

III. FILTERING STAGES

After collecting predicted sequences from the two class algorithms and verified them by the one class algorithms finally we sorted out 571 potential pre-miRNAs from 1439 pre- miRNAs by using a cutoff, which is the predicted pre-miRNA has to be selected by minimum of two different algorithm from chrosome 3 and 4 . Then these 571 sequences are passed to the following filtering stages.

A. Cross species conservation search

MiRNAs are conserved across species; to verify this we have used BLAST 2 alignment algorithm from NCBI. The individual pre-miRNA sequences are aligned to genome sequences of Plasmodium falciparum and Plasmodium knowlesi. We have set a cutoff that is 17 nt of predicted pre miRNA sequences should have a 100% match with the genome sequence of either P.falciparum or P. knowlesi. The other cutoffs are expect thresholds- 10, word size- 11, match/mismatch- 2,-3, Gap costs- existence5, extension2, and filtering the low complexity region.

B. Coding sequence search

MiRNAs are non-coding; to verify this we have used nucleotide BLAST algorithm of NCBI with reference mRNA sequence (refseq_rna) database. The hypothetical mRNA matches have been inferred as non coding sequences alone with the no match sequences. The cutoffs are just like the Blast2 except word size- 28, match/mismatch- 1,-2 and gap costs- Linear.

As per n-Blast result, this sequence is having 100% matches with Plasmodium vivax Sal-1 calcium-dependent protein kinase coding sequence. Therefore, this sequence is not a pre-miRNA sequence.

C. Structural filter for the following properties

Structural filter; for this purpose, we have used a tool named RNA analyzer. It predicts the following features: Sm site or snRNP site; if yes the probability of having exon portion in test sequence or there is a presence of catalytic RNA GG-pairs; if yes Influence of mis-pairing on DNA backbone conformation. Two or three stem structure: if yes probably a t-RNA structure. Protein A1 binding site; if yes the probability of having apoptotic regulatory protein binding site. AU rich region; if yes, an AU-rich element or "ARE" is a region with frequent A and U bases in an mRNA that targets it for degradation. Cleavage stimulation factor binding region (CstF); if yes probability of a pre-mRNA sequences.

IV. MIRBASE SIMILARITY SEARCH

After filtering stages, 239 potential pre-miRNAs are compared with miRNA database. For this, individual sequences are aligned with the miRBase. We have chosen the stem loop sequence from search sequence option alone with search method BLASTN. The e-value cutoff sated at default 10. All the sequences having e-value <=1.0 are finally selected as most possible putative pre- miRNAs. There are 153 pre-miRNAs are sort listed. The flow diagram of this process has been shown in appendix 1. According to this result, the input sequence found match with E-value 0.25. So, this sequence is considered as pre- miRNA sequence.

V. RESULTS

After analyzing chromosomes no. 3 & 4 of Plasmodium vivax genome by six algorithms a table has been prepared. A small part of the combined table 2 has shown below: The combined table: 3 consists of chromosome number, specific codes for the selected sequences by MirEval and ProMiR 2, the respective positions and separate columns for algorithms. The different colour represents number of the algorithms like, orange coloured sequences were selected by minimum and maximum two algorithms, the green coloured sequences were selected by minimum three and maximum 4 algorithms and the red coloured sequences were selected by minimum five and maximum six algorithms. These coloured sequences have been taken to the next level of filtering procedures. Up to now 1439 predicted pre-miRNA sequences have been verified by all six algorithms. The 1439 pre-miRNA sequences consist of chromosome number 3 and 4. We have sorted out 571 predicted pre-miRNA sequences for the filtering stages from these 1439 predicted pre miRNA sequences by a cutoff; every pre- miRNA sequence should be selected by minimum of two algorithms.

TABLE 2 Describing the list of unfiltered putative pre-miRNAs and their different parameters

Project miRNA: Comperism Chart. (ORENGE- min 2-max2; GREEN- min 3-max4; RED- min 5-max6)

No.	Chro No.	MirEval Code	MirEval Position	Triplet SVM	Struc. Cluster	ProMir Code	ProMir Position
1	1	Me1	4041 to 4120	Y	N	N	N
2		Me2	4161 to 4240	Y	N	N	N
3		Me3	4581 to 4670	Y	N	N	N
4		Me4	5661 to 5740	Y	Y	N	N
5		N	N	N	N	X1	6211 to 6310
6		Me5	7091 to 7170	Y	N	N	N
7		Me6	8661 to 8740	Y	N	N	N
8		Me7	10281 to 10380	Y	Y	N	N
9		Me8	10771 to 10850	Y	N	N	N
10		Me9	12341 to 12430	Y	N	N	N
11		Me10	13871 to 13980	Y	Y	X2	13891 to 13990
12		Me11	16671 to 16760	Y	N	N	N
13		Me12	19361 to 19440	Y	N	N	N
14		Me13	21041 to 21130	Y	N	N	N
15		Me14	21431 to 21510	Y	N	N	N
16		Me15	23861 to 23950	Y	N	N	N
17		Me16	24531 to 24610	Y	N	N	N
18		Me17	25211 to 25290	Y	N	N	N
19		Me18	25451 to 25540	Y	N	N	N
20		Me19	25631 to 25710	Y	Y	N	N
21		Me20	26741 to 26820	Y	N	N	N
22		Me21	28531 to 28610	Y	N	N	N
23		Me22	28801 to 28880	Y	N	N	N
24		Me23	30471 to 30550	Y	N	N	N
25		Me24	32211 to 32290	Y	N	N	N
26		Me25	32951 to 33030	Y	N	N	N
27		Me26	34721 to 34800	Y	N	N	N
28		Me27	35191 to 35280	Y	N	N	N
29		Me28	35861 to 35940	Y	Y	N	N
30		Me29	36441 to 36520	Y	N	N	N
31		Me30	38121 to 38200	Y	N	N	N
32		Me31	39471 to 39560	Y	Y	N	N
33		Me32	40751 to 40830	Y	N	N	N
34		Me33	42641 to 42730	Y	N	N	N

The filtering stages and the procedures have previously mentioned. After filtering information we have created another table. A small portion of combine table 4 is presented below.

TABLE 3. Selected list of miRNAs after phase 1

Project MicroRNA: Chart of selected MiRNA from Phase 1

No.	Chro No.	MirEval Code	ProMir Code	Sequence	Combine Position
1	1	Me4		ATCATACCCCACT	5661 to 5740
2		Me7		TTGCCGAGTAATT	10281 to 10380
3		Me10	X2	GCAAAGGTAGGAA	13871 to 13990
4		Me19		CATACGCGTTCCT	25631 to 25710
5		Me28		GGATAACGAATTG	35861 to 35940
6		Me31		AGTCATTCTGTGTG	39471 to 39560
7		Me34		ATTGGTAAATGTC	42781 to 42860
8		Me35		TGGGTAGTGGACA	44291 to 44380
9		Me38		GCGAGGGGGTCAA	45351 to 45470
10		Me42	X3	TATTCCAGAGGTTC	50921 to 51020
11		Me43		GAACAATGCCCTT	51561 to 51650
12		Me44		AAATGGAAGAAGT	51811 to 51950
13		Me45		GGGCCAAGTGGCG	52971 to 53050
14		Me49	X4	GTGCGCACACACG	64051 to 64220
15		Me50		CGATATGCAGCGC	64541 to 64620
16		Me51		TACTGAAGTACTG	66841 to 66930
17		Me52		CGCATGATGAAGA	67781 to 67870
18		Me56		TAACACACATGTG	72171 to 72250
19		Me64		CCCACCGGTTGGA	84681 to 84790
20		Me65		ACACGTTGGTACG	86091 to 86170
21		Me67	X5	GTTGGCTAGCTCG	86921 to 87040
22		Me68	X6	TGCGGCCATCACG	87751 to 87880
23		Me72		GGGAGCTCCACAT	89871 to 10000
24		Me73		GTGTGTGCTCTTTC	91091 to 91280
25		Me74		GTTCATCCACGTC	91921 to 92010
26		Me78	X7	GGCACTCCCACTG	96251 to 96350
27		Me80		GGAAGTGGGATAT	99211 to 99320
28		Me82		ACAATTTCTCCATA	100121 to 100280
29		Me83		AAACGCTTCCGCC	100811 to 100940
30		Me88		GAACGAACGGTGT	113121 to 113280
31		Me97		TTTAAGCGAAAGT	126501 to 126660
32		Me98		GTGCTTCGCCAAA	127011 to 127090
33		Me99		CAACAAAGCAGTT	127361 to 127490

TABLE 4. Partial table showing selected candidates after filtering.

In table 4, the filtering type 1 consists of BLAST 2 results of Plasmodium falciparum and Plasmodium knowlesi. The conserved sequences are denoted as "Y" with the chromosome number. Other sequences are denoted as "N".

No.	Chro No.	MirEval Code	ProMir Code	Combine Position	miRBase Search Results
1	3	Me10	X2	13871 to 13990	Y
2	3	Me28	No	35861 to 35940	Y
3	3	Me42	X3	50921 to 51020	Y
4	3	Me49	X4	64051 to 64220	Y
5	3	Me52	No	67781 to 67870	Y
6	3	Me67	X5	86921 to 87040	Y
7	3	Me68	X6	87751 to 87880	EV=1.3
8	3	Me105	no	134081 to 134170	Y
9	3	Me123	X8	145201 to 145300	Y
10	3	Me170	X12	200021 to 200120	Y

The filtering type 2 consists of nucleotide BLAST results and the filtering type 3 consists of RNA analyzer results. Till now we have tested 571 pre- miRNA sequences by the three filtering approaches. From 571 pre miRNA sequences we have filtered out 239 pre- miRNA sequences by considering the following point: all the conserved sequences are considered as pre-miRNA if they are non-coding and if they are not snRNP motif or 2-3 Stem structure. Because snRNP motif is the splisosome binding site and the 2-3 stem structures are consider as tRNA sequence. Then the 239 pre-miRNAs are aligned with miRBase and candidates having e-value (EV) <=1.0 are finally selected as pre miRNAs. The result in this stage is shown in the combined table 5.

TABLE 5. Showing partial list of finally predicted putative miRNAs

No.	Chromosome No.	Combine Position
1	3	13871 to 13990
2	3	35861 to 35940
3	3	42781 to 42860
4	3	50921 to 51020
5	3	51811 to 51950
6	3	64051 to 64220
7	3	66841 to 66930
8	3	67781 to 67870
9	3	72171 to 72250
10	3	86921 to 87040

There are 153 sequences are selected as putative pre-miRNAs. In the combine table 4, miRBase results are shown for every sequence. The sequence having EV <=1.0 are denoted as "Y". The entire predicted 153 pre- miRNAs are shown in appendix 2.

VI. DISCUSSION

There are many well known tools available for the miRNA prediction. But we used this two and one class combinations of six algorithms simply because the lack of knowledge about miRNAs from Plasmodium vivax. There is no such tool specific for this genera or species. We have used triplet SVM because of its ability to reduce the rate of false positives along with the structure clustering algorithm to introduce more structural parameters for pre-miRNA prediction. After that we have used proMiR 2-g for finding of both the clustered and non clustered pre miRNA sequences. Some of the parameters of these algorithms are some what same like stem loop free energy, conservation search etc. While doing the analysis of the predicted data we have seen the different number of predicted sequences by different tools or algorithms. Like 1439 pre- miRNA sequences has been predicted from chromosome number 3 and 4, by triplet SVM algorithm of MirEval. Among them ~571 sequences has been selected by structure clustering algorithm. On the other hand only 76 pre- miRNA sequences have been predicted by ProMiR 2. This vital difference in number of outputs is due to the prediction parameters. We have selected the other species option in MirEval. Because of this the system (triplet SVM) treats the inputs by general trained properties and user undefined cutoffs. Then it provides selected sequences when ever it finds matches like stem loop structures and etc. But in case of other algorithm it selects the sequences with the known cluster matches and by using structural parameter which are mainly found in terns of whether or not the sequence recognized by RNase 3 enzymes like Drosha and Dicer. On the other side in ProMiR 2, we have selected C. elegance from the species selection box. Because of this system try to find out the similar conserved sequences and cluster matches with the known data of C. elegance. This is the reason behind the huge number difference between algorithms. So there can be a high number of false positive sequences present in this 1439 predicted pre- miRNA sequences. On the other side ProMiR 2 should be missing many true positive sequences because of Plasmodium vivax has an undetermined negative class. To solve this problem we use another tool which is learned by only the positive class. We use three different algorithms to select the most possible pre- miRNA sequences. We have verified individual results by these three algorithms (a) OC-SVM, (b) OC-Gaussian, (c) OC-KNN. At first we have decided to take sequences which have been selected by minimum of four algorithms but in that case we have to lose some data.

So finally we have taken the cutoff of only two algorithms and 571 sequences are taken to the filtering stages. In the filtering stages we have selected three types of filtering. All of them are dealing with the basic properties of miRNA. MicroRNAs are conserver between species. For that we have used primary sequence conservation search by BLAST 2 between predicted sequences and genome sequences of P. f and P.k and we have found few highly conserved sequences. But many sequences are not selected as conserved. After that to find out the non coding sequences from the predicted pre – miRNAs we have used the n BLAST program against the known mRNA database. By this two filtering stages we have eliminated all the coding sequences and the non conserved sequences which are also coding sequence. But there are few structural properties need to be clarified like splisosome binding sites and tRNA sequences. The tRNA have stem loop structure but they contain more then one or two loops and the splisosome binding sites are very near to the exon and sometimes with in an exon. Though there is evidence of miRNAs within the exons but we avoid this property to make our study simple. To solve the above mentioned problems we have used RNA analyzer. It finds the tRNA s and splisosome binding sites. Finally we have sorted out 284 pre- miRNA sequences from the chromosome number 3 and 4. These 284 pre miRNAs are aligned with the miRBase sequences for similarity search. 153 pre miRNAs find matches with the database sequences. So we calculate 63% similarity with our final set of data. This may be because of the lack of data from Plasmodium genera. The other chromosomes will be tested soon and added to this data. The selected candidate pre miRNAs will be verified experimentally and we will reach our goal by finding out their targets in human genome. There is also a need of parasite or Plasmodium genus specific miRNA prediction tool for more accurate findings.

REFERENCES

[1] Vasudevan S, Tong Y, Steitz JA (2007) Switching from repression to activation: microRNAs can up-regulate translation. Science 318:1931–1934

[2] Ruvkun G (October 2001). "Molecular biology: Glimpses of a tiny RNA world". Science (journal) 294 (5543): 797–9. doi:10.1126/science.1066315.

[3] World Malaria Report 2008, by World Health Organization.

[4] William Ritchie1,2,*, Franc¸ois-Xavier The´ odule1 and Daniel Gautheret (2008); MirEval: a web tool for simple microRNA prediction in genome sequences; Vol. 24 no. 11 2008, pages 1394–1396 doi:10.1093/bioinformatics/btn137

[5] Chenghai Xue†2,1, Fei Li†1, Tao He1, Guo-Ping Liu2,3, Yanda Li1 and Xuegong Zhang*1 (2005); Classification of real and pseudo microRNA precursors using local structure-sequence features and upport vector machine; BMC Bioinformatics 2005, 6:310 doi:10.1186/1471-2105-6-310

[6] William Ritchie,1 Matthieu Legendre,1,3 and Daniel Gautheret2 (2007); RNA stem–loops: To be or not to be cleaved by RNAse III; RNA. 2007 April; 13(4): 457–462. doi: 10.1261/rna.366507.

[7] Jin-Wu Nam1,2, Jinhan Kim3, Sung-Kyu Kim1,2 and Byoung-Tak Zhang (2006); ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs; Nucleic Acids Research, 2006, Vol. 34, doi:10.1093/nar/gkl321

[8] Malik Yousef1,3, Segun Jung1,2,4, Louise C Showe1 and Michael K Showe (2008); Learning from positive examples when the negative class is undeterminedmicroRNA gene identification; Algorithms for Molecular Biology 2008, 3:2 doi:10.1186/1748-7188-3-2

[9] Leaman, D., Chen, P. Y., Fak, J., Yalcin, A., Pearce, M., Unnerstall, U.,Marks, D. S., Sander, C., Tuschl, T. and Gaul, U. (2005). Antisense-mediated depletion reveals essential and specific functions of microRNAs in Drosophila development. Cell 121,1097 -1108.

[10] Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. (2003) Identification of Drosophila MicroRNA targets. PLoS Biol 1: E60.

[11] Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E. et al. (2004) A pancreatic isletspecific microRNA regulates insulin secretion. Nature 432: 226–230.

[12] Jin P, Alisch RS, Warren ST (2004) RNA and microRNAs in fragile X mental retardation. Nat Cell Biol 6: 1048–1053

[13] Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci USA. 2002; 99(24): 15524-9.

[14] Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, Iorio MV, Visone R, Sever NI, Fabbri M, Iuliano R, Palumbo T, Pichiorri F, Roldo C, Garzon R, Sevignani C, Rassenti L, Alder H, Volinia S, Liu CG, Kipps TJ, Negrini M, Croce CM. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med. 2005; 353(17): 1793-801.

[15] Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with Antisense complementarity to lin-14. Cell 75, 843–854.

[16] Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21 nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403, 901–906.

[17] Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294, 858–862.

[18] Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science 294, 862–864.

[19] Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science 294, 853–858

[20] Bartel, B. and D.P. Bartel. 2003. MicroRNAs: At the Root of Plant Development?

[21] Lim L.P., M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. 2003. Vertebrate microRNA genes.

[22] Cullen, Bryan R. (2006). Viruses and microRNAs. Nature Genetics 38, S25-S30.

[23] Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., Ruvkun, G. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408(6808): 86-89.

[24] Michel J. Weber New human and mouse microRNA genes found by homology search, FEBS Journal Volume 272, Issue 1, pages 59–73, January 2005

[25] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M and Schuster P. (1994) Fast Folding and comparison of RNA secondary structures. Monastch. Chem., 125:167-188