International Journal of Computer Theory and Engineering, Vol. 2, No. 2 April, 2010
1793-8201

Prediction of miRNAs in Bombyx mori through Computational Approaches

A.K. Mishra, D.K. Lobiyal

Abstract— **In this paper, we have used rule-based approach to predict mature miRNA from pre-miRNAs of Bombyx moori (Silk worm). For rule based approach, it is important to have a set of rules. Unfortunately, to the best of our knowledge, miRNA related literature does not provide well-defined set of rules for miRNA prediction. Therefore, the challenge in this research was to derive a set of rules from known data set of related species. We have made an attempt to derive a set of rules from the known miRNA and pre-miRNA data sets of Drosophila melanogaster and Apis mellifera. The derivation of rules is based on the dominating features identified using PCA and infogain method applied by authors in their previous work. Based on these rule set a PERL script is written to predict mature miRNA from given pre-miRNA sequences of Bombyx mori. The results are good approximation with a small number of mismatches. The results of the research are encouraging and may facilitate miRNA biogenesis.**

Index Terms--**miRNA, Prediction, Rule based approach, Sequences.**

I. Introduction

MicroRNAs (miRNAs) are an evolutionary conserved class of non coding RNAs of small length approximately 20 to 25 nucleotides (nt) long and found in diverse organisms like animal, plant etc. miRNAs play very important role in various biological processes. They regulate gene expression at post- transcriptional level by repressing or inactivating target genes [1, 2]. miRNA biogenesis is highly associated with stem-loop feature of its precursor's secondary structure. As pre-miRNA secondary structures consisting of stem-loop are highly conserved across different species, extracting informative attributes from secondary structure is significant step in identification of miRNA from unknown sequences [3,4]. The biochemical-based methodology used for identification of novel miRNAs, in the laboratories can be assisted by computational methods. Computational methods can identify various potential miRNA that can further be verified by the former approach. Therefore, researchers have developed computational models for miRNA prediction. miRNA gene finding is a challenging aspect in field of computational biology. It is due to the fact that non-coding family of RNA does not posses any strong statistical signal as well as it lacks generalized algorithms.

This is evident from various studies that hairpin structure of miRNA is dominating feature in the secondary structure and quite informative in retrieving and inferring biological information. Eukaryotic genome contains high number of inverted repeats, which on transcription converted into hairpin structure. Due to its large number in the genome it has become quite important to choose the right hairpin among them and this is a major problem suffered by biologists. Computational methods have given attempt to reduce the search space and enhance accuracy, which therefore proved helpful in searching the right hairpin for prediction [5,6].

There are different approaches to predict the secondary structure of miRNA. These can be classified as energy minimization based, grammar based, matching based and evolutionary algorithm based approaches. Free energy minimization is one of the most popular methods for the prediction of secondary structure of RNA. Energy minimization methods use the dynamic programming approach along with some sophisticated energy rules. The energy of the predicted RNA structure is estimated by summing negative base stack energy of each base pair and by adding positive energy of destabilizing regions like hairpin, bulges and internal loops [7,8]. A number of tools have been developed based on energy minimization method and widely used by computational biologists.

The approaches to detect miRNA genes can be categorized into 3 types: forward genetics, direct cloning and computational approaches using bioinformatics tools.

The first miRNA genes, lin-4 and let-7 were discovered by forward genetics method [9,10]. This method involves the identification of mutation responsible for a certain phenotype. But this approach has not been used very much to detect miRNA families due to certain limitations like small size of miRNA and the 'seed' sequence, which doesn't mutate easily [11]. On a large scale miRNAs were identified using direct cloning method. In direct cloning, small RNA were isolated from biological samples and cloned to make a cDNA library for small endogenous RNAs [12]. This approach is limited only for highly expressed miRNAs. Low expressed miRNAs are hard to clone. So some miRNAs may be missed due to such sequence-based biases.

The limitations of above two methods can be overcome by the use of bioinformatics approaches and genome sequence analysis. Bioinformatics approaches mainly rely on the information about certain miRNA properties. This information is based on the known miRNA features set generated by cloning method. The set of distinct structural features of miRNA makes the base of their computational prediction. Secondary structure information of pre-miRNA is used by most of the methods. The prediction involves first the searching of miRNA precursor and then detection of mature miRNA from it. The computational approaches to identify mature miRNA from its precursor can be categorized into 3

Manuscript received August 14, 2009
1Jawaharlal Nehru University, New Delhi, India
2U.S.I, Indian Agricultural Research Institute, New Delhi, India

types:(1) Rule based approach; (2) Machine learning approach; (3) Homology based approach [13].

In a rule based approach, some common features to most of the miRNA are identified manually. These features help to define some rules for the detection of miRNA genes. A single rule is not enough for the accurate prediction of miRNA genes. However, if some rules are used collectively, prediction can be made effectively. In a machine learning approach, the features are trained using some known datasets by automated procedures. Homology based approach uses the highly conservative nature of miRNA among species. Some rely only on the primary structure information while some methods use both primary and secondary structure information for the phylogenetic conservation.

Research community has been using different algorithms for identifying potential miRNA in a precursor. miRscan and miRseeker are two computational tools for identifying miRNA from precursor. miRscan[14,15] uses a sliding window of size 21 nucleotides that passes through a precursor for locating probable miRNA. Scores for every window is computed based on the features set derived from phylogenetically known miRNA. miRseeker is a secondary structure alignment based tool with certain degree of divergence among the structures. It aligns secondary structures of precursors where one is for known miRNA and other in which miRNA is being identified.

MiRScan starts with two sequences of two closely related genomes. MiRScan assigns a log-likelihood score to each position of the sequence on the basis of seven features by sliding a 21 nt window over the hairpin. Then it aligns it with other sequence to identify probable miRNA gene.

The modified version of this tool came in existence in the form of MIRSCAN II [15a], which contains certain additional features like conserved motifs etc. It has been proved beneficial in finding motifs in different species like Homo sapiens and *Drosophila melanogaster*. miRseeker [16] was used as a three part computational pipeline to identify miRNA genes in *Drosophila*. The initial step was the identification of conserved region of D. melanogaster and D. pseudoobscura genome using multiple alignment technique. This was followed by the identification of conserved stem-loop using Mfold and then evaluating them on the basis of pattern of nucleotide divergence characteristics of known miRNAs. It was also estimated that *Drosophila* genome contains more than 150 miRNA genes. It was a good approximation because till date only 157 miRNa is available in Sanger's repository.

A tool called miRFinder [17] has been developed to predict pre-miRNA that is conserved between two genomes. It uses 18 attributes for miRNA prediction. This program consists of three major steps. First an algorithm based on the Smith Waterman algorithm which scans the genome pair wise sequence to extract the high potential regions to form a hairpin. The criteria for selection were hairpin length should be greater than 18 nucleotides and multiple loops not allowed.

Another approach based on the study of Wang et al. [17a] for the prediction of probable *Arabidopsis thaliana* miRNA genes. This approach used characteristic features of miRNA and sequence conservation between *Arabidopsis thaliana* and *Oryza* sativa genomes. Using this approach 83 novel miRNA were found. Out of these 83 novel miRNA experimental verification was done for 25 miRNAs. The computational pipeline filters the miRNA and their precursors on the basis of certain features like stem-loop structure, GC content of the mature miRNA sequences, number and distribution of mismatches in the hairpin etc.

II. Motivation

miRNA and pre-miRNA prediction is still not a widely explored research area in the domain of computational biology. However, researchers have been using limited techniques for miRNA prediction as mentioned in the previous section. Some of the prediction models are developed using different sets of features of miRNA from model organisms. We also focused on exploring dominating set of features from *Apis mellifera* using Principal Component Analysis (PCA) and Infogain in our earlier work with the assumption that it will reduce complexity of the model and may increase prediction accuracy. The results of previous works from both the methods were encouraging and required to be used for prediction of miRNA to prove the assumption of the earlier work. Further, in the literature rule based approach using features is not widely explored by the researchers. In this paper we have explored rule-based approach for miRNA prediction in *Bombyx mori*.

III. Rules Collected From Literature

Rule-based miRNA detection typically looks at the minimum free energy (mfe) structure in pre-miRNA. Researchers have identified common rules from most known pre- miRNA of different species. Some of the major rules are summarized as follows:

- No more than 1 of the 20 nt may be asymmetrically unpaired [18].
- At least 16 of the bases in the mature miRNA have to be paired [19].
- GC content must be ≥ 0.3 and ≤ 0.7 and with an entropy value ≥ 1.75 [20].
- miRNAs have a lower folding free energy than random sequences (non miRNA) of the same nucleotide composition [20].
- miRNA::miRNA* duplex matches were restricted to ≤ 7 mismatches, ≤ 3 continuous mismatches, and ≤ 2 gaps in the 25 nucleotides centered on the miRNAs and miRNA*s in accordance with published plant miRNA annotation guidelines [21].
- MFEs normalized by candidate miRNA precursor length were required to be below −0.3 kcal/mol· bp, and were derived from the aforementioned analysis of the known miRNA [21].
- The candidate miRNA hairpin precursor loop length was required to be greater than 15 bp given that the minimum known miRNA precursor length in the RNA registry is about 55 bp [22].

- Vertebrate organisms have a conserved U at the beginning of the sequence and a GU rich region in the end around position 18–25[23].
- In case of plants 5' arm motif having a strongly conserved U and the 3' arm a conserved C at the mature end [23].

IV. Derived Rules

The rules collected from literature are fuzzy in nature and varies from species to species. Therefore, miRNA prediction with these rules is not viable and needs further exploration of better rules. We have derived important rules based on the available datasets in hexapoda family. For this purpose we have systematically analysed pre-miRNA datasets of *Apis mellifera,* Droshophila *Melanogaster,*Anophole *gambie* etc. These rules play very important role in mature miRNA identification. The rules we derived are stated below:

- Not more than 1 of the 20 nt are asymmetrically unpaired
- At least 16 of the bases in the mature miRNA has to be paired
- GC content must be >= 0.3 and <= 0.7
- The pairing must extend 4nt beyond the mature miRNA
- There is a bias in occurrence of first five bases in a sequence (especially a U at the first position)
- a tendency toward having symmetric rather than asymmetric internal loops and bulges in the miRNA region
- Presence of 4-6 base pairs between the miRNA and the terminal loop region
- There is a high probability of occurrence of miRNA and miRNA* near the hairpin. The probability of miRNA occurrence decreases as we move away from the hairpin.

Some of the relationship observed between miRNA and miRNA* is as under

- ≤ 7 mismatches
- ≤ 3 continuous mismatches
- ≤ 2 gaps in the nucleotides on miRNAs and miRNA*s
- MFE < −0.3 kcal/mol·bp,

V. Material And Methods

The work reported in this paper has been carried out in following five phases:

- Data collection
- Structure prediction
- Attribute measurement
- Attribute reduction based on Information gain
- Implementation of derived rules using PERL script

A. miRNA Precursor Data Collection

There are very limited open and free domain sources of miRNA data available to the research community. miRBase is one of the highly referred database easily accessible for miRNA research. We downloaded the complete dataset of 62 known pre-miRNA of *Apis mellifera* (honey bee) and 156 known pre-miRNA of *Drosophila Melanogaster* from miRBase sequence database (a data repository of published microRNA sequences and its annotation) (release 11.0) at http://microrna.sanger.ac.uk [24,25]. Further, 62 and 156 non pre-miRNA sequences were taken from *Apis mellifera* and *Drosophila Melanogaster* genome data respectively on random basis..

B. Predicting miRNA Secondary Structure

In this phase, choice of a secondary structure prediction tool was important. First we decided to use an energy based structure prediction model available in open and free domain. We selected RNAfold over Mfold since it runs on windows and gives its output in a form for which a program was developed for further manipulation of the structure. RNAfold software is small windows based utility of Vienna RNA secondary structure server available at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi RNAfold produces a single structure that has minimum free energy for a given sequence. Therefore, we produced secondary structures for the each entity both pre-miRNA and non pre-miRNA of *Apis mellifera* and *Drosophila melanogaster*..

C. Measuring Attributes from pre-miRNA Secondary Structure

In this phase, the values of 14 attributes from 124 secondary structures were measured. This resulted in a relation with 116 tuples of 14 dimensions each. First, we define each of the 14 elements from the following attribute set used in our work:

$$S_A = \{LEN, NBP, BLR, NHP, HPL, FE, FEN, AUC, MSK, SD, MBL, MBS, MTL, NTL\}$$

LEN attribute represents the length of the input pre-miRNA sequence. NBP is number of base pairs. BLR denotes the ratio of base pair to sequence length. NHP is used for number of hairpins. HPL denotes average length of hair pins in a structure. FE is free energy in K cal/mol of a optimal structure. FEN is ratio of Free energy to length for normalizing free energy to sequence length. AUC denotes % of AU contents in the sequence. MSK denotes maximum number of continuous base pairs. SD is symmetric difference, i.e. it is difference between the lengths of both arms of the structure. MBL is length of the bulge of maximum size. MBS is maximum number of bulges in an arm either upper or lower). MTL denotes maximum tail length. NTL is number of tails in the structure of the given sequence.

We developed our program in C++ that reads the following output of the phase (b) as an input and produces the values of all 14 attributes. The secondary structure in dot

bracket form for *ame-mir-7 MI0001594* sequence is as follows.

(((((((.(((...(((.((.(((((((((((((((((.....)).)))))).)))))))))..)))..)))..)))).)))))))

In this output the brackets represent base pairing and dots denote mismatches in the precursors of miRNA sequences.

D. Attribute reduction and classification

We have used Weka software (version 3.5.8) [26] for Principal Component Analysis (PCA) and Information gain evaluation of the dataset for dimension reduction and relevance analysis. Weka takes input in a specific format. Therefore, the data obtained after measuring the values of attributes is transformed into the Weka input format (.arff). After applying PCA using Weka on the dataset the ranked attributes corresponding to eigenvectors were obtained. These are ranked in the order of amount of variation in dataset. We selected top five attributes with variance more than 0.1. Finally by dividing attribute weights by the variance of five eigenvectors the calculated values for the attributes in eigenvectors are ranked. Attributes BLR, HPL, and AUC in fourth eigenvector (V4) has highest values 1.729, 1.474, 1.189, respectively. Attribute FEN in second eigenvector (V2) has its value 1.536. The highest value of other attributes in any eigenvector is less than 1.0. Therefore, we selected these four attributes as dominating

We have applied best first search method in forward direction to determine the essential/dominating attributes. Based on information gain four essential attributes were determined namely LEN, NBP, FEN and AUC. We have computed decision tree based on essential attributes obtained from information gain analysis. We have produced a tree-based classifier using j48 (a weka implementation of C4.5 algorithm) for pre-miRNA classification [27,28].

E. Implementation of derived rules using PERL script

The domination or essential attributes are derived using PCA and Infogain methods are described in section (d). Based on these attributes and critically analyzing the data from *Apis* Mellifera and Drosophila Melanogaster some salient rules were derived. The details are discussed in section IV. We have assigned different weights for each rule The rules were implemented after assigning scores to each rule using PERL script to derive miRNA from pre-miRNAs of *Bombyx moori.*

VI. Results And Discussion

First, we introduce a sample of dataset with fourteen attribute values extracted from pre-miRNA database thorough our C++ program. The attribute values extracted from the secondary structure (given in figure 1) of *ame-mir-7 MI0001594* sequence

GAGCGCCGUUGCAUGGAAGACUAGUGAUUUUGUUGUUCUACUUUCGAUAUAACAAGGAAUCACUAAUCAUCCUACAAAGGCGCUCG are as follows in the order they appear in attribute set *A*. The secondary structure of this in bracket and dot form is already given in section (c) of material and methods.

{87, 33, 0.37, 1, 7, -37.20, 0.42, 56, 17, 2, 3, 6, 0, 0}

We have derived relevant attributes namely FEPN, NBP, AUC and LEN using PCA and information gain analysis on known dataset of pre-miRNA. Further a decision tree was derived for the training dataset.

```
          g   cau   a  c          --       -  ua
 cgagcgcc uug   gga ga uagugauuu  uguugu uc  c
 |||||||| |||   ||| || |||||||||  ||||||  ||  u
 gcucgcgg aac   ccu cu aucacuaag  acaaua ag  u
          a   -au   a  a          ga      u  cu
```

Fig. 1 Secondary structure of *ame-mir-7 MI0001594*

We have tested our model on synthesized data sets of 500 and 100 instances. The model has worked perfectly on small set. Above results clearly show that our model gives high precision and recall in all cases. However, the model is not able to recall few true cases this may be improved by taking a larger training dataset. The large training dataset may also help in identifying dominating and essential attributes for increasing the accuracy of prediction model.

Applying these dominating attributes in training datasets of *Apis Mellifera*, *Anopholes gambiae* and *Drosophila Melanogaster* some salient rules were derived, which is also justified by biologists miRNA detection rules. We have assigned scores based on the rules and fixed the threshold to 110. Above the score of 110 we have considered the miRNA as suitable candidate. Based on these scoring criteria we are able to predict 48 miRNAs in *Bombyx mori* using a script written in PERL.

Predicted miRNA of *Bombyx moori* is given as under*:*

Predicted miRNA	Score
CUGAGAUCAUUGUGAAAGCUAA	125.09
UGCGACUGUAUAGCCUGCU	142.69
AGUAUGGAGCUGCGCGGGC	158.15
GACAAAUUCGGUUCUAGAGAGG	134.70
GUAUUACUUCAGGUACUGGUUG	111.46
CACUGCCGGAGCCGUUAUG	139.43
GGUCCCCUUCAACCAGCUG	131.64
GGUCACGCGUAUUCUUGGG	132.16
CGAUCCUGUCAAAGCGGCGGUG	150.59
GGCCCACUCGUAAAAAUGGCUGUG	146.42
GGGAGAGAAAUCGACGAGGCUG	139.25
GCAGUGACGUGCCCUUGUC	153.22
GCCGCGCGUCGUCGAGAUAUG	143.44
ACCUCUCUGGCGCGCGCGU	153.22
CGACUUCCGUUCCUGGCGGGG	135.70
CCAGCAGCUCCUUCCCGAGCG	145.74
CCCGAUGUUGCUUGACUUCGG	123.87
GAAUCCAAACGCUUUGCCC	121.64
GCUGCUGGCCACUGCACAAG	150.07
CCCGUGAUCUCUUAGUGGC	123.42
CCGUGAAUUUCCCGAUGCC	136.91
AGGCGGCAGCGCCGCGCGC	173.94
GUAGGAACUUCAUACCGUGC	125.59
CCUCGGGGUUCGUGCCAGG	169.01

GACGAACUUCCCAGCUCGGCC	140.68
GAGUGGAGGUUUAGUGCAUG	132.89
GAGAGCUAUCCGUCGACAG	136.91
CCCAGGCUAUCAGCUGGUA	147.43
CGAAUUCAGUGCUCGAACGUGG	127.34
GCCAGCUUUGAUGAGCACGAC	138.16
GCCAGCUUUGUUGAGUUCUCGGC	148.41
GUUCCAGGCGCUUGUUGGAG	142.37
CUCACUCAACCUGGGUGUG	123.42
GAGCCGGUGGCUGGGAAGGC	150.07
CGGGUGCCACGCUGUGCUC	152.70
GGCAGUGUGGUUAGCUGGU	142.17
CACUAAUCUGCCUACAAAGCG	119.63
GAAAGACAUGGGUAGUGAGA	142.37
GGCGCGCUGCGACGCUUUG	152.70
GCAUCUUACCGGGCAGCAUUAGAG	144.03
CAGGUGUGUUAGUGCCGGC	142.17
ACGUCAUAAAGCUAGGUUACCGG	128.64
CCCGGCCUGCCUGUGGCCAC	145.07
GGGCAAAGCGUUUGGAUUC	118.37
CGCGACUCCCUAAUCGAGUC	137.89
GCGGGAGUGAGGCUGAGGC	147.95
CAAAGCGUCGCAGCGCGCC	158.48
CGCCUUCUGUACGUAUACUG	135.59

VII. Conclusion

In this paper we have derived pre-miRNA attributes from their secondary structure and used PCA and Information gain approaches for selecting essential attributes based on their relevance. The results are encouraging since the essential attributes selected here are biologically significant. These attributes are used in deriving rules for miRNA identification. Based on these rules we are able to identify miRNAs of *Bombyx mori* with a fair accuracy by allowing few mismatches. For other species the rules can be further modified and may be applied for prediction of miRNAs of other related species. Performance of this model can be enhanced by incorporating other rule based mining techniques and scoring criteria. This forms the future direction of our research.

Acknowledgment

The authors are thankful to Jawaharlal Nehru University, New Delhi (INDIA) for providing support through capacity build-up funds.

References

[1] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75: 843-854

[2] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. Cell 115: 787-798

[3] Bartel,D.P.:MicroRNAs:Genomics,biogenesis,mechanism and function.Cell116(2004)281–297

[4] Lee,Y. et al. The nuclear RNaseIII Drosha initiates microRNA processing. Nature425(2003)415–419

[5] Tinoco Jr., I., Uhlenbeck, O.C., and Levine, M.D. (1971): Estimation of secondary structure in ribonucleic acids. Nature 230: 362–367

[6] Zuker, M. and Stiegler, P. (1981): Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9: 133–148.

[7] Nussinov, R. and Jacobson, A.B. (1980): Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl. Acad. Sci. 77: 6309–6313.

[8] Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999): Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288: 911–940.

[9] Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. (2002): MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 21, 4663–4670

[10] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000): The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403:901–6

[11] Eugene Berezikov, Edwin Cuppen & Ronald H A Plasterk (2006): Approaches to microRNA discovery. Nat. Genet.,38,june2006

[12] Ambros, V. and Lee, R.C. (2004): Identification of microRNAs and other tiny non-coding RNAs by cDNA cloning. MethodsMol. Biol. 265, 131–158

[13] Morten Lindow and Jan Gorodkin (2007): Principles and Limitations of computational Microrna Gene and Target Finding .DNA & Cell Biology 26

[14] Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S.,Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003a). The micro-RNAs of Caenorhabditis elegans. Genes Dev. 17, 991–1008.

[15] [15] Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P (2003b). Vertebrate microRNA genes. Science 299, 1540.

[16] Ohler,U., Yekta,S., Lim,L.P., Bartel,D.P. and Burge,C.B. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA, 10, 1309–1322.

[17] Lai EC, Tomancak P,Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. Genome Biology.2003;4(7):R42

[18] Huang TH, Fan B, Hu ZL, Li K, Zhao SH (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans BMC Bioinformatics

[19] Wang XJ, Reyes JL, Chua NH, Gaasterland T. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biology. 2004; 5(9):R65.

[20] M. Jones-Rhoades, D. Bartel Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA Molecular Cell, Volume 14, Issue 6, Pages 787-799 2005

[21] Lindow and Krogh, Principles and Limitations of Computational MicroRNA Gene and Target Finding 2005 DNA AND CELL BIOLOGY Volume 26, Number 5, 2007

[22] Eugene Berezikov* and Ronald H.A. Plasterk Camels and zebrafish, viruses and cancer: a microRNA update Human Molecular Genetics, 2005, Vol. 14, Review Issue 2 R183–R190 2005

[23] Jones-Rhoades, M. W., Bartel, D. P. and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. Annu. Rev. Plant Biol. 57, 19-53).

[24] Jiayu Wen, Tancred Frickey and Georg F. Weiller; Computational prediction of candidate miRNAs and their targets from Medicago truncatula non-protein-coding transcripts

[25] J. Gorodkin et al. MicroRNA sequence motifs reveal asymmetry between the stem arms, Comput. Biol. Chem. 30 (2006) 249–254

[26] Griffiths-Jones S, Saini Hn Dongen S, Enright AJ miRBastools for microRNA genomics NAR 2008 36(database issue) D154-D-158

[27] Griffiths-Jones S, Grocock RJ, Van Dongen SBatchan A, Enright AJ miRBase: microRNA 2006 sequencesetsand gene nomenclature 34(Database issue) D140-D144

[28] Fra E. et al. mining in bioinformatics using Weka Bioinformatics 20(15) 2004

[29] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[30] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996

A.K.Mishra completed his Master degree in Computer Application from Kumaon University Nainital (India) in 1996 He is currently working as a Scientist with Unit of Simulation & Informatics, IARI, New Delhi (India) and a Ph.D candidate from JNU, New Delhi. His area of interest is

Bioinformatics and Computer Application in agriculture. He is a professional member of IACSIT (Singapore) and CSI (India).

D.K.Lobiyal completed his Ph.D in Computer Science from JNU, New Delhi (India) in 1996. He is currently working as Associate Professor in School of Computer & Systems Sciences JNU, New Delhi (India). His area of interest is Mobile Ad-Hoc Network, Bioinformatics & Natural language processing. He is fellow of IETE (India).