International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010
1793-8201

Naïve Bayes Classifier for Classification of Plant and Animal miRNA

Bhasker Pant, Kumud Pant and K. R. Pardasani

Abstract— **MicroRNAs (miRNA) are single-stranded RNA molecules of about 21–23 nucleotides in length. MicroRNAs (miRNAs) constitute a large family of non coding RNAs that function to regulate gene expression. Till today wet lab experiments have been used to classify the miRNA of plants and animals. The wet lab techniques are highly expensive, labour intensive and time consuming. Thus there arises a need for computational approach for classification of plants and animal miRNA. These computational approaches are fast and economical as compared to wet lab techniques. In view of above a machine learning models has been developed for classification of plant and animal miRNA using Naive Bayes classifier. The model has been tested on available data and it gives results with 85.71% accuracy.**

Key words— **Micro RNA, RNA interference, Naïve Bayes Classifier.**

I. Introduction

We want to understand the basic process of miRNA functioning because they have wide application in development of biotech products, diagnostics, drug development, agro industry and therapeutics ex. miRNA based drugs [5]. Animal miRNAs have been found to play a major role in diseases like cancer, heart diseases, neurological disorders and aging. The miRNAs have been found to play specific roles in plant development, including the regulation of flowering time and floral organ identity, and leaf polarity and morphology.

They can play a major role in Agro Biology for enhancing crop productivity and also increasing resistance towards major pests and diseases.

There is a growing interest in identifying miRNA and determining their role in skeletal muscle and adipose tissue development in cattle. Also the miR motifs are associated with feed efficiency which is an important factor that represents greater than 50% of the total cost in most livestock production systems [6]. Hence correct identification of miRNA that regulate cellular processes and impact economically important traits is the need of the industry.

Classification developed here is useful in evolution studies where miRNA belonging to particular plant specie show conservation with other species this shows evolutionary decent. They have application in forensic science where miRNA belonging to organism can be identified and as the classification is extended further incorporating all organisms in the mirBASE registry more specific analysis can be done. The miRNA classified can be shown to have relationship with the sequence, structure and function of the genes lying nearby. The upstream and downstream genomic region can be identified with miRNA classification and signature [1].

The miRNA can have association with various molecular characteristics and structures leading to new function. Their association with mRNA can describe the role they play in the organism hence the resultant function of genes they control can be useful in systems biology.

It has been shown in the recent studies that the structure of miRNA precursor stem–loops exhibits a significantly high level of mutational robustness in comparison with random RNA sequences with similar stem–loop structures. This is not due to base composition bias or thermodynamic stability but it is the result of direct evolutionary pressure towards increased robustness [20]. This requires better understanding of characteristics of miRNAs which can be done by understanding the differences between miRNAs of different organisms.

Various attempts have been made till date to discover novel miRNAs in various species of plants and animals by using both in-vivo and in-silico techniques and elucidate their role in various regulatory processes. But from literature survey it appears that no attempt has been made to develop computational approaches for classification of plant and animal miRNAs, thus there is a need to develop newer algorithms which are robust, fast and economical considering the financial and time constraint which it poses on existing lab techniques.

There are many similarities between plant and animal miRNA system, both system play fundamental role in development and appear to predominantly exert their influence by controlling regulatory genes. Using computational techniques we can identify an object, in this case a gene as belonging to a particular class. For the classification to be successful, each class must show some distinct properties or characteristics.

In both plants and animals miRNAs post transcriptionally regulate gene expression through interaction with their target mRNAs and these targets are often genes involved with regulating key developmental events [11]. Despite these similarities, plants and animal miRNAs exert their control in fundamentally different ways. In animals the first step of miRNA biogenesis involves Drosha, but this role is performed by DCL1 in plants [1], [15]. The majority of plants miRNAs are each derived from single primary transcripts from loci found in the intergenic regions, whereas many of animal miRNAs are generated from polycistronic transcripts from intergenic regions of the chromosomes and many are produced from introns. In plants miRNAs mainly regulate their targets by cleavage in the coding regions of the RNA

Manuscript received 16th September 2009.

B. Pant is with Bioinformatics Department, MANIT, Bhopal, India (e-mail: pantbhaskar2@gmail.com).

K. Pant is with Bioinformatics Department, MANIT, Bhopal, India (e-mail: pant.kumud@gmail.com).

K. R. Pardasani is with Department of Mathematics, MANIT, Bhopal, India (email: kamalraj@gmail.com).

whereas animal miRNAs mainly operate by translation repression using targets at the 3'-UTR [12], [13]. But there is almost always an exception that breaks the rule. None of the characteristics described in the above sections allows by itself a direct classification of a given miRNA as plant or animal miRNA. This work, however, tests if multiple features can be combined to create a powerful classifier. The algorithm of choice is the Naive (or simple) Bayes classifier that finds its origins in the Bayesian theory of probability. In present work we have developed a naïve bayes classifier for classification of the plant mirna and animal mirna, using the differences between the plant and the animal mirna which are given in the Table 1

TABLE1. FEATURES OF PLANT AND ANIMAL MIRNAS

	Plants	Animals
Number of miRNA genes present	100-200	100-500
Location within genome	Predominantly intergenic region	Intergenic region intron
Presence of miRNA clusters	Uncommon	Common
miRNA biosynthesis	Dicer like	Drosha ,Dicer
Location of miRNA binding motifs within target gnes	Predominantly the open reading frame	Predominantly the 3'-UTR
Number of miRNA binding sites within target genes	Generally one	Generally multiple
Function of known target genes	Regulatory genes crucial for development enzymes	Regulatory gens crucial for development structural protein ,enzymes

II.MATERIALS AND METHOD

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or lack of presence) of a particular feature of a class is unrelated to the presence (or lack of presence) of any other feature depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods [2], [3].

The main advantage of Bayesian classifiers is that they are probabilistic models, robust to real data noise and missing values. The Naive Bayes classifier assumes independence of the attributes used in classification but it has been tested on several artificial and real data sets, showing good performances even when strong attribute dependences are present. In addition, the Naive Bayes classifier can outperform other powerful classifiers when the sample size is small [4], [7]. Since it also has advantages in terms of simplicity, learning speed, classification speed, storage space and incrementality its use should perhaps be considered more often.

A model is developed to classify Animal and Plant miRNA on the basis of physical characteristics only. It exploits features already available in databases, like number of miRNA genes present, presence of miRNA in clusters, number of binding sites and complementariness and combines them into a Naive Bayes classifier to obtain the classification of animal and plant miRNA. These features are given in Table 1.

A NB classifier calculates the probability that a given instance (example) belongs to a certain class. It makes the simplifying assumption that the features constituting the instance are conditionally independent given the class. Given an example X, described by its feature vector (x1,...,xn), we are looking for a class C that maximizes the likelihood. We consider each data instance to be an n-dimensional vector of attribute values:

X = (x1, x2, x3. . . xn).

After data filtering, each classifier is trained and cross-validated for 10-times with a 10-fold random sampling. The ten resulting values for each performance parameter are averaged to obtain the final figures and Receiver Operating Characteristic (ROC) curves and TP rates vs. FP rates are plotted and analyzed. As expected, the Naive Bayes classifier shows a definite progression in performance when data discretisation is used, either supervised or un-supervised.

A.Software

The miRNA target registry software is used to extract the properties of the animal and plant miRNA. The Weka Data Mining Java script 3.6 was used for training and testing the Naive Bayes classifier [14].

B.Classifier

Previously miRNAs have been classified using decision tree classifier, the algorithm for which was also taken from Weka suite. The same dataset has been used here also [21]. All the algorithms used were taken from the Weka suite [10].

C.Evaluation

The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use stratified 10 –fold cross validation. All evaluation parameters are calculated with a ten times, tenfold cross-evaluation. The method uses nine tenths of the data for training the system while the remaining tenth is set aside as a test set (control) for estimating the various evaluation parameters, like the success rate (see below for parameters definition). The data is randomized and the procedure is repeated 10 times to estimate the average value for each parameter [8], [16] – [18]. Extensive tests on numerous datasets with different learning techniques have shown that 10 is about the right number of folds to get the best estimate of error. The parameters used for evaluation are the following (where TP = true positive, FP = false positive, TN = true negative and FN = false negative)

[19].

D.Precision

Defined as the number of positive instances retrieved over the total number of instances declared positive by the classifier =TP/TP+FP.

E.Recall

It is defined as the number of true positive instances retrieved over the total number of instances that are positive in the set =TP/TP+FN.

F.F Measure

Combines precision and recall = 2TP/2TP + FP + FN.

G.Success Rate

It is the number of real positive and negative instances retrieved over the total number of instances.
= TP + TN / TP + TN + FP + FN

H.Root Mean Squared Error

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

Where p1, p2... pn are the predicted values for each transcript, a1, a2... a_n are the actual values and n is the total number of predictions (number of transcripts considered). The standard deviation over the ten success rate values and over the ten root mean squared error values is calculated as follows.

$$\sqrt{\frac{\sum_{i=0}^{N} (x_i - \bar{x})^2}{N-1}}.$$

Graphical description of various attributes used in the classifier is shown in figure 1, 2, 3 & 4. Each and every attribute is checked with the given class and their frequency is plotted in the form of graph.

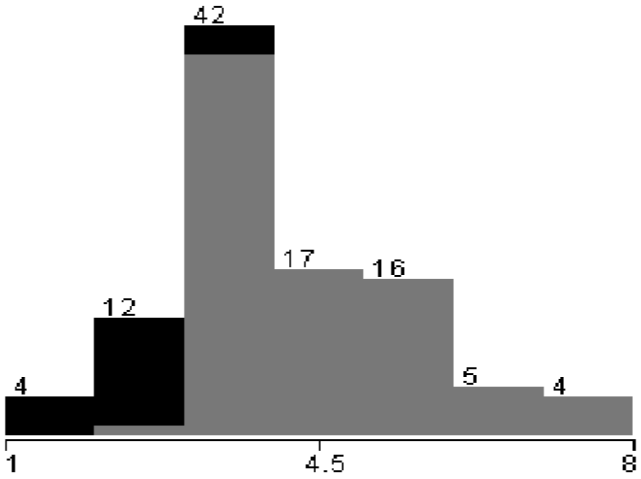


Fig. 1 Number of mismatches with target mRNA (on X axis)/ vs. Number of miRNA (on Y axis).

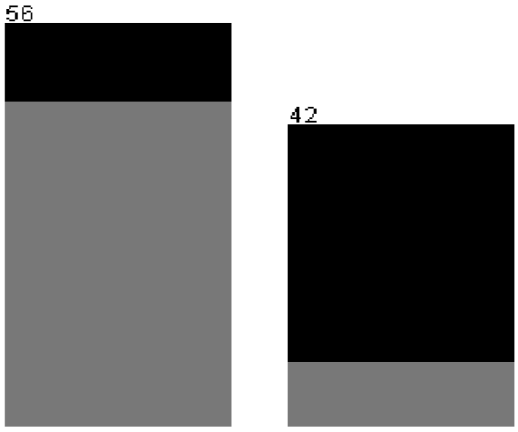


Fig. 2 Complementarity of miRN (on X axis)/ vs. Number of miRNA (on Y axis).

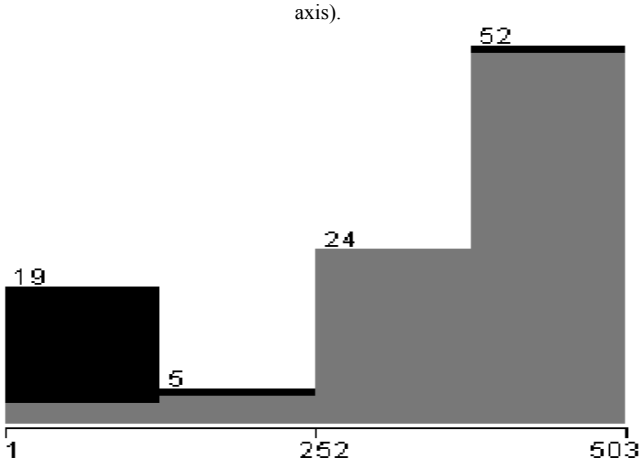


Fig. 3 Number of target genes (on X axis)/ vs. Number of miRNA (on Y axis).

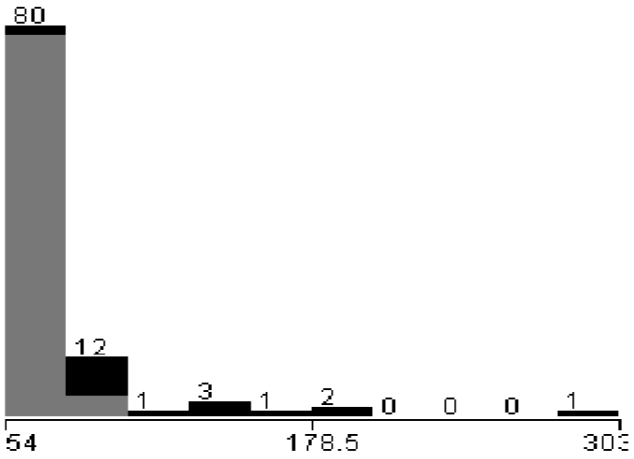
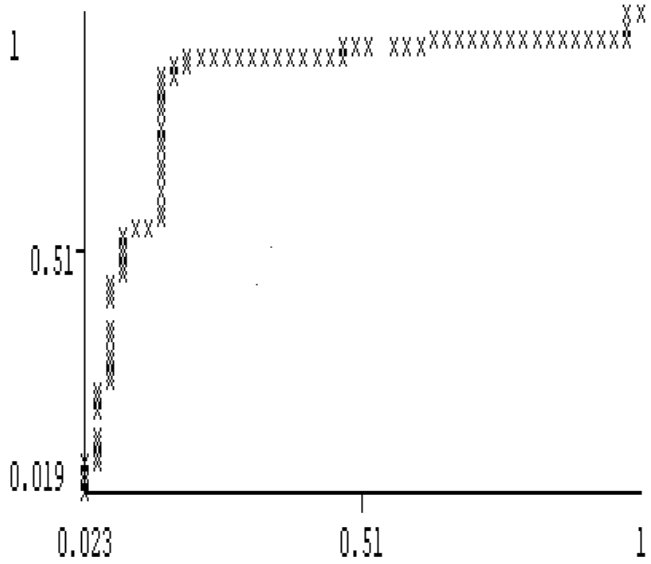


Fig. 4 Size of fold back loop (on X axis)/ vs. Number of miRNA (on Y axis).

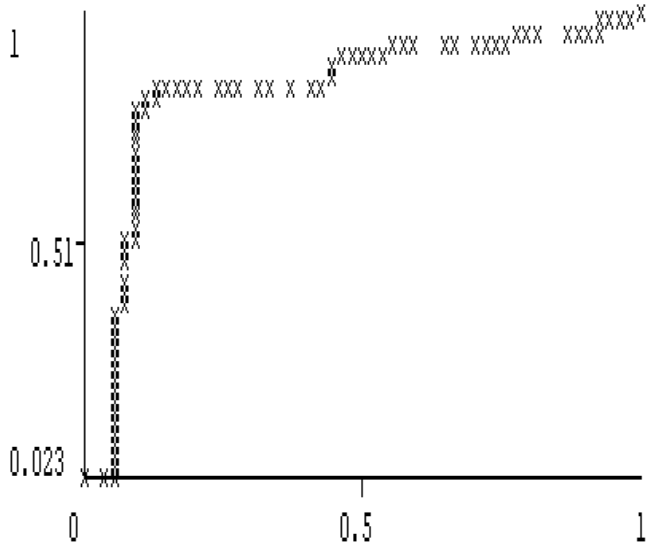
I.Training Set

For the classification of animal and plant miRNA we select the dissimilarity between the animal and plant miRNA. For the classification purposes we have selected number of mismatches with target mRNA. The number of mismatches in plants have values less than or equal to 3 but for animals this value is 4 or more than 4 which implies that plant have

greater . Complementarity of miRNA with target has higher value for plant and low for animals. Size of fold back loop is greater than 100 for plant with variations till 303 nucleotide and less than 100 nucleotides for animals. With these characteristics we have trained the weka classifier and the values we get are given in the table 2.



Graph1-ROC Curve for animals



Graph2-ROC Curve for plants

III. Results And Discussion

A set of attributes was collected and the corresponding attribute values were fed to the classifier for each transcript of all plant and animal miRNA genes. Not all attributes, however, are fit for use in a classifier [9]. The attributes like presence of cluster are the attributes not taken for classification purpose since the information was not adequate. The result show 85.71% classified instances and 14.29% unclassified instances .The detailed characteristics of the classification is given in table 2 and 3 which uses naïve bayes classfier.

A.Figure 1

Shows the number of mismatches with mRNA. The black portion of the graph indicates data for plants, majority of which are found to have zero or less then or equal to three mismatches, very few have four. For animals the grey graphs indicate number of mismatches which can be as many as 8. None of the miRNAs in animals have 0 mismatches.

B.Figure 2

Show complementarity of miRNA with target mRNAs. Majority of plants have high complementarity, whereas animals have low complementarity with target mRNA.

C.Figure 3

Shows that number of target genes which in case of plants is less, whereas for animals the number of target genes exceeds as much as 503. Very few plant miRNAs have large number of target genes.

D.Figure 4

Shows the size of fold back loop. For animal the variation in loop is lesser whereas for plants the loop size varies as much as 303 nucleotides.

E.ROC curves

They depict the performance of a classifier without regard to class distribution or error costs. The horizontal axis represents false positive and vertical axis represents true positive. The value of ROC curve for this data set comes out to be 0.835 which represents these values in the true positive region as shown in graph1 and graph2.

TABLE2. Stratified Cross Validation (Summary)

Correctly classified instances	84	85.7193%
Incorrectly classified instances	14	14.2857%
Kappa statistics	0.7101	
Mean absolute error	0.1658	
Root mean squared error	0.3679	
Relative absolute error	33.4728%	
Root relative squared error	73.901%	
Total number of instances	98	

TABLE3. Detailed Accuracy by Class

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.889	0.182	0.857	0.889	0.873	0.835	Animal
0.818	0.111	0.857	0.818	0.837	0.835	Plant
0.857	0.15	0.857	0.857	0.857	0.835	-Weighted average

Confusion Matrix

```
 a   b   <-- classified as
48   6 | a = Animal
 8  36 | b = Plant
```

IV. Conclusion

There are many obvious similarities between plant and animal miRNA systems; both systems play fundamental roles

in development and appear to predominantly exert their influence by controlling regulatory genes. However, there are also many differences. Since this is a probabilistic model hence it can only be validated if more and more results are tested and checked. The accuracy and reliability of the model depends upon the amount and quality of data input to the model. Exceptional cases can also be checked and verified using the above approach and further validation can be achieved through wet lab experiments. Using the probabilistic combination new insights are provided for wet lab experiments.

We have been developed a computational approach to discover if a miRNA is animal or plant using simple features of that miRNA and its surroundings. This is made possible by the integration of different attributes operated by the Naive Bayes engine, which are available in literature. The classifier developed in this paper can be used to generate information which may l be helpful in understanding the inherent properties of miRNAs in both plant and animal systems, formulate various association rules, clustering algorithms and other data mining activities which the author intends to carry out in future.

Here we observe that four characteristics (complementarity, number of mismatches with target mRNA, number of target genes and size of fold back loop) used in the model to develop the classifier give us fairly good accuracy in results ie 85.71% accuracy. Here we infer that these four characteristics are very important and must be included in any classifier of plant and animal miRNA's. Apart from these there are characteristics which were dependent and their inclusion in the model does not bring any improvement in performance and accuracy of the model. Other characteristics like size of miRNA family and number of binding sites within target genes etc can be included in the classification models to improve the performance and efficiency of the classifier but the limitation is that sufficient information is not available about these additional characteristics in the literature at present. However as soon as sufficient information about more additional characteristics becomes available in the literature, the authors intend to include the same in the above classifier in future to improve its performance and accuracy.

ACKNOWLEDGEMENT

The authors would like to thank Department of Biotechnology New Delhi, India and M.P. Council of Science and Technology, M.P., India for Bioinformatics infrastructure facility.

REFERENCES

[1] A. Anthony Millar, M.Peter Waterhouse, "Plant and animal microRNAs: similarities and differences," SpringerLink Funct Integr Genomics, 2005, 5: 129–135.
[2] CSE5230 Tutorial: The Na¨ıve Bayes Classifier 1.
[3] P. Baldi, "Bioinformatics the Machine Learning Approach ," Soren Brunak, 2nd Edition, 2001.
[4] R.O. Duda, P.E.Hart and D.G.Stork, "Pattern Classification," John Wiley, 2nd edition, 2001.
[5] L. Aagaard and John J. Rossi, "RNAi Therapeutics: Principles, Prospects and Challenges," Elsevier Science, 2007.
[6] T.G. McDaneld, R.T. Wiedmann, J.R. Miles, R. Cushman, J Vallet and T.P.L. Smith, "USDA/ARS U.S. Meat Animal Research Center, Clay Center, NE MicroRNA technology in livestock: expression profiling of bovine oocyte and developmental stages of porcine skeletal muscle," 2007.
[7] .Luna De Ferrari and Stuart Aitken, "Mininghousekeeping genes with a Naive Bayes classifier School of Informatics, the University of Edinburgh, Edinburgh EH8 9LE, UK," 2006.
[8] IH. Witten, E. Frank, "Data Mining – Practical machine learning tools and techniques with Java implementations," Morgan Kaufmann, San Francisco, 2005.
[9] De Ferrari L, "Mining housekeeping genes with a Naïve **B**ayes classifier University of Edinburgh (MSc Thesis)," 2005.
[10] Weka Data Mining Java Software, [http://www.cs.waikato.ac.nz/~ml/weka/].
[11] M.W. Jones-Rhoades, D.P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," Mol Cell, 2004, 14:787–799.
[12] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B.Burge, B. Bartel, D. P. Bartel, "Prediction of plant microRNA targets," Cell, 2002, 110:513–520.
[13] E. M. Meyerowitz, "Plants compared to animals: the broadest comparative study of development," Science, 2002, 295:1482–1485.
[14] Micro RNA Registry, www.microrna.sanger.ac.uk/
[15] V. Ambros, "The functions of animal microRNAs,"15. Nature, 2004, 431:244–350.
[16] P. Langley and S. Sage, "Elements of machine learning," San Fracisco: Morgan Kaufmann, (1994).
[17] J. Han and M. Kamber, "Data mining: concepts and techniques," San Francisco:Morgan Kaufmann, 2001.
[18] D. Heckerman, D.Geiger and D. M. Chickering, "Learning Bayesian network: The combination of knowledge and statistical data. Machine Learning," 1995, 20(3):197-243.
[19] R.C. Holte, "Very simple classification rules perform well on most commonly used datasets," Machine Learning, 1993, 11:63-91.
[20] E. Borenstein and E. Ruppin, "Direct evolution of genetic robustness in microRNA," School of Computer Science and School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel, 2006.
[21] B. Pant, K. Pant and K. R. Pardasani. "Decision tree classifier for classification of Plant and Animal microRNA's," SpringerLink Communication in Computer and Information Science, 2009, 51: 443-451.

B. Pant received his M. Sc. degree in Computer Science from Barkatullah University, Bhopal, M. P., India and C-DAC from Silverline institute for software technologies, Chennai, Tamil Nadu, India. He is currently pursuing Ph.D degree under the guidance of Dr. Kamal Raj Pardasani in Bioinformatics at Maulana Azad National Institute of Technology (MANIT), Bhopal, India.

K. Pant received her M. Sc. degree in Botechnology from Jiwaji University Gwalior, M. P., India. She is currently pursuing Ph.D degree in Department of Bioinformatics at Maulana Azad National Institute of Technology (MANIT), Bhopal, India.

K. R. Pardasani is currently working as Professor and Head Department of Mathematics, Bioinformatics and Computer Application at Maulana Azad National Institute of Technology (MANIT), Bhopal, India. He has more then 150 research publications in both National and International Journals. A keen researcher and able administrator with multifaceted personality.