Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 67(1) 2013 59-70

Attribute Analysis and Classification of miRNAs

A.K. Mishra[1] and D.K. Lobiyal[2]
[1]U.S.I., Indian Agricultural Research Institute, New Delhi
[2]Jawaharlal Nehru University, New Delhi

Received 14 November 2011; Revised 21 February 2012; Accepted 26 February 2012

SUMMARY

Several models for miRNA prediction based on attributes of known miRNAs have been developed. These models are based on different sets of attributes. However, only limited attempts have been made in exploring dominating attributes to reduce the complexity of the model and increase the classification accuracy. To the best of our knowledge statistical techniques for attribute selection have not been applied for miRNA attribute analysis. miRNA prediction, generally consider attributes from few model organisms widely used by researchers from biological sciences. Further, most of the models focus more on training algorithms to improve the prediction accuracy rather than attributes relevance analysis. In this paper we have derived 14 attributes for precursors, 9 attributes for mature miRNAs and 20 attributes in combination from both precursors and mature miRNA for relevance analysis and classification of four hexapode species-*Apis mellifera, Bombyx moori, Anopheles gambiae and Drosophila Melanogaster*. Dominating feature extraction was done using different machine learning techniques from a set of known miRNA sequences for the above mentioned species using PCA, Infogain, SVM attribute analysis, Cfs subset, Consistency subset evaluation and Chi squared analysis. The results are encouraging since the essential attributes selected here are biologically significant. These attributes can be used in deriving rules for miRNA identification. The performance measures on training and test datasets are quite satisfactory. The results obtained from experiments clearly show that our model gives high precision and recall for all the four species in all combinations. Future developments ought to focus on the need to establish more accurate models using sophisticated algorithms of artificial intelligence techniques and rule based mining approach.

Keywords: miRNA, Attribute, Classification, Relevance, Prediction.

1. 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 (Lee *et al*. 1993, Lewis *et al*. 2003). Primary miRNA (pri-miRNA) is several hundred nucleotide transcript processed in the nucleus by Drosha which is a multi protein complex containing enzymes (Lee *et al.* 2003). It generates a ~70-nt long in case of animals and ~60 to a few hundred nucleotides in case of plants miRNA stem–loop precursor (pre-miRNA). The secondary structure plays a vital role for Drosha substrate recognition rather than the primary sequence (Cullen 2004).

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 (Bartel 2004, Lee *et al*. 2003).

Corresponding author : A.K. Mishra
E-mail address : akmishra@iari.res.in

There are different approaches to predict the secondary structure of RNA. 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 (Tinoco *et al*. 1971, Mathews *et al*. 1999).

2. CLASSIFICATION AND RELEVANT ATTRIBUTES

Classification is to put objects in one of a pre-specified set of categories or classes. Identification of common characteristics of a set of objects that are representative of their class is very useful in many computational biology applications. For example, to classify whether a given sequence fold to a stem-loop structure or not and similarly to classify whether a given sequence is miRNA or not. In machine learning, classification is commonly dealt using supervised learning. In supervised learning there is a specified set of classes, and objects labeled with appropriate class. The aim is to identify the class of new objects based on the knowledge obtained from training data set. The success of classification depends on the quality of the data used to train the classifier (Witten and Frank 1999).

To build a simple model with acceptable level of performance using training data set a small number of features (attributes) is preferred over a large number to reduce complexity of the model. The presence of irrelevant and redundant attributes in the dataset makes the model unreliable and noisy. Therefore, learning with training data becomes more difficult. Attribute relevance analysis is the process of identifying and removing irrelevant and redundant attributes from the dataset.

Identification and removal of irrelevant attributes from datasets is essential before they are used to build the models. Models to evaluate the relevance of attributes can be grouped in two categories – Wrapper and filters. The Wrapper evaluates datasets by using a learning algorithm, while the filters use heuristics for evaluation (Kohavi 1995, Kohavi and John 1996).

Conventionally, researchers use regression analysis to determine the contribution of a factor in making a decision. There are two limitations of such approaches. First, application of the statistical methods is based on the assumptions that variables under study are independent of each other. Second, model is represented in the form of coefficients, which require mathematical expertise. Therefore, model based on machine learning algorithms are becoming more popular, since they do not require data mining expertise in understanding the classifier.

Attribute Reduction and Relevance Analysis Techniques

Biological data is mostly very large in size and redundant. This characteristic of data makes it difficult to find algorithms that are time and space efficient. Therefore, it becomes important to reduce the redundancy and dimension of the datasets. In this section a brief discussion on attribute reduction and attribute relevance analysis techniques used in bioinformatics domain is given.

Principal Component Analysis (PCA)

PCA is used for transformation of data from higher dimensionality to lower. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of variance in the original data. Noise in attributes can be filtered by transforming the data to the principal component space by eliminating some of the worst eigenvectors and then transforming back to its original space.

Let a set of attributes $A = \{a_1, a_2, \dots a_n\}$, each attribute represents some characteristics features of pre-miRNA such as hairpin, bulges, base pairs, etc. PCA uses correlation between (a_i, a_j) to generate a correlation matrix C of $n \times n$. The eigen values of C are computed by solving the following characteristics equation for C to obtain a set of orthogonal eigenvectors.

$$[I\lambda - C] = 0 \qquad (1)$$

The eigen values $\lambda_1, \lambda_2, \dots \lambda_m$ where $m \le n$ results into $E_v = \{e_1, e_2 \dots, e_m\}$ where e_i is an eigen vector and $m < n$. Further E_v space is transformed into reduced space $E_r = \{e_1, e_2 \dots, e_l)$ where E_r is set of top raking eigenvectors $l \le m$, and $E_r \subset E_v$. Each eigenvector represent an attribute weight vector. It is assumed that attributes with large weights contribute more to the principal component than the attributes with lower weights. Finally using both the weights of $a_i \in A$ and the amount of variance by the principal components, $A = \{a_1, a_2, ..., a_n\}$ transformed into $D = \{b_1, b_2, ..., b_k\}$ where, $b_i \in A$, $k < n$, and D represents the set of dominating attributes and $D \subset A$.

Information Gain

Entropy based Information gain approach is used for dimensionality reduction and attribute relevance analysis. For this, it is assumed that X be a set of training samples, where class label of each sample is either 'True' or 'False' denoting whether the sample is pre-miRNA or not. There are two classes, therefore, X contain x_1 samples of miRNA and x_2 samples of non miRNA with probability x_1/x, where x is the total number of samples in set X. The expected information needed to classify a given sample is

$$I(X_1, X_2) = \sum_{i=1}^{2} \frac{x_i}{x} \log_2 \frac{x_i}{x} \tag{2}$$

An attribute A with values $\{a_1, a_2, ..., a_v\}$ can be used to partition X into the subsets $X_1, X_2, \ldots, X_v$ where X_j contains those samples in X that have value a_j of A. Let X_j contain x_{ij} samples of class Y_i. The expected information based on this partitioning is known as the entropy of A. Entropy is the weighted average computed as given below:

$$E(A) = \sum_{j=1}^{v} \frac{x_{1j} + x_{2j}}{x} I\ (X_{1j}, X_{2j}) \tag{3}$$

The information gain obtained by partitioning using A is defined by

$$\text{Gain } (A) = I(X_1, X_2) - E(A) \tag{4}$$

For every attribute A from S_A information gain, Gain (A_i) is obtained to rank the attributes based on the decreasing value of information gain. Using ranking of the attributes, relevant attributes are selected. These relevant attributes are used to construct decision tree based classifier that is used to predict pre-miRNAs and miRNAs.

Subset Evaluation using Cfs

It evaluates the relevance of a subset of attributes by considering the individual predictive ability of each attribute and the degree of redundancy among attributes. Subsets of attributes that are highly correlated with the class, and having low inter-correlation among them self are preferred (Hall 1999).

Subset Consistency Evaluation

Subset consistency evaluation is filter based method of attribute selection. It evaluates the worthiness of a subset of attributes by the degree of consistency in the class when the training instances are projected on the subset of attributes. This evaluator is used in conjunction with a random or exhaustive search (Hall 1999).

Attribute Evaluation using SVM

This method uses a linear support vector machine to determine the value of attributes using recursive feature elimination technique. It evaluates the worth of an attribute by using an SVM classifier. Attributes are selected one by one based on the size of their coefficients. A fixed number of attributes can be removed at each stage. This process continues till a certain number of attributes remain and then each attribute is analyzed intensively.

Attribute Evaluation using Chi-Square

It evaluates the dominance of an attribute by computing the value of chi-square statistic with respect to the class. This evaluation method uses ranker based search for evaluation of attributes.

3. CLASSIFICATION TECHNIQUES

In this section a brief discussion on decision tree based classification along with other popular and widely used classification techniques is given.

Decision Tree

Decision Tree (DT) is one of the most popular choices for learning from feature-based examples. They are especially attractive in data mining environment for several reasons. First, due to their intuitive representation, the resulting classification model is easy to assimilate by humans. Second, DT does not require any parameter setting from the user and therefore, are specially suited for exploratory knowledge discovery. Third DT can be constructed relatively faster and its accuracy is comparable or superior to other classification models.

DT is a hierarchical structure with root node at the top and the leaves at the bottom. Following a top down approach, DT is constructed recursively by splitting the given set of examples. DT uses two algorithms-ID3 for discrete attributes and its successor C4.5 for numeric continuous attributes. A DT may be mapped to a set of decision rules by traversing all the paths from the root towards leaves (Quinlan 1986).

C4.5 algorithm summarizes training data in the form of a decision tree. Due to its robustness and execution speed decision tree algorithms are very popular in biological data mining community. Nodes in the DT correspond to attribute (feature), and branches to associated values of attributes. The leaves of the tree correspond to classes. To classify a new instance, one can simply examine the attributes at the nodes of the tree and follows the branches corresponding to their observed values in the instance. After reaching a leaf, the process terminates and the class at the leaf is assigned to the instance (Quinlan 1987, 1989).

Other Classification Techniques

Genetic algorithm, Neural Network (NN), Fuzzy logic, Bayesian network and Instance base learning are some of the popular classification techniques in data mining. Genetic algorithms are inspired by Darwin's theory of evolution. A population of possible solution to a problem is initially created on random basis. These initial solutions are used to generate more feasible solution by using genetic operators, until an optimal solution is obtained. Genetic algorithms are applied for a variety of bioinformatics applications, such as the discovery of patterns in sequences. Neural network (NN) is an approach based on biological neurons for developing mathematical models with the ability to learn. NN models are well suited for prediction and forecasting problems. Fuzzy logic is derived from fuzzy set theory dealing with approximate reasoning rather than, reasoning precisely deduced from classical predicate logic. Conceptually, the fuzzy truth represents the membership in a vaguely defined set. A Bayesian Network (BN) is a model for probability relationships among a set of features. The most interesting feature of BN is consideration of prior information about a given problem, in terms of structural relationships among its features. Instance-based learning algorithms delay the induction or generalization process until classification is performed. It requires less computation time during the training phase but more computation time during the classification process.

4. ATTRIBUTE RELEVANCE ANALYSIS FOR miRNA

The work reported in this chapter for miRNA attribute relevance analysis has been carried out in following four phases:

- Data collection
- Structure prediction
- Attribute measurement
- Attribute reduction and relevance analysis

miRNA Data Collection

As mentioned earlier miRBase is one of the highly referred databases easily accessible for miRNA research. A dataset of 54 known pre-miRNA of *Apis mellifera* from miRBase (release 11.0) was downloaded. This repository contained only 54 known pre-miRNA of *Apis mellifera*. Further, 54 non pre-miRNA sequences were taken from *Apis mellifera* genome data. Complete dataset of 91 known miRNA sequences of *Bombyx mori*, 157 known miRNA sequences form *Drosophila melanogaster* and 66 known miRNA sequences from *Anopholes gambiae* were taken from miRBase (release 14.0). Further, equal number of non-pre-miRNA sequences was taken from genome data randomly to make negative dataset.

Predicting RNA Secondary Structure

In this phase, choice of a secondary structure prediction tool was important. First we compared grammar based and energy based models. From grammar-based models, we have selected Contrafold. In energy based models, RNAfold and RNAstructure were selected. Contrafold and RNAfold give secondary structure in two-dimensional representation as well as in dot bracket form. RNAstructure gives secondary structure output in two-dimensional representation only. We have converted the two-dimensional output to dot bracket form for our ease using crimson editor and a C++ script.

For example for a given *bombyx mori* sequence

AUCGGUGGUGAAGAGAAGGUGUAUGGAGGAUG UGAAGGUCCAGAUGCUAUGUAUGUGAAACUGG

UAUCCUCUGAUGGACAUGAGUUCAUUGUGAAGAGGG AGCAUGCUCUUAUAUCAGGCACUAUUAAGGCUAU GCUCAGCGGACCGGGCCAGUUUGCUGAAAA UGAAGCUAACG secondary structure using these three tools is as under.

................((((..((((((((......((((.((.......))....))))))))))))((((((....))))))..((...((((()))))))...))..)))).....((((
((..(((((((((.......)))))))))..))..)))).... Contrafold

...((((((......(((((.((((.((((...(((.(((.((((...)))).))))..))....(((((((((((....))))))....)))))).))))).))))))))...))))))..(((((..
(((((((((.......)))))))))..))..)))).... (–48.90 Kcal/mol.) RNAfold

...((((((......(((((.((((.((((....((((.((((((.((......))))))))))(((((((((((....))))))....))))))))))).))))))))...))))))..(((((.
.(((((((((.......)))))))))..))..)))).... (–48.4 Kcal/mol) RNA structure

We decided to use an energy based structure prediction model available in open and free domain. These models provide minimum free energy to fold the sequences besides secondary structure. We selected RNAfold over RNAstructure 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 produces a single structure that has minimum free energy for a given sequence. Therefore, we produced secondary structures for each entity both for pre-miRNAs and non-pre-miRNAs of hexapods.

Measuring Attributes from pre-miRNA Secondary Structure

In this phase, the values of 14 attributes from all instances of pre-miRNA were measured. First, we define an attribute set S_A of 14 attributes used in our work:

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

Each attribute in the set S_A discussed below represent some characteristics of pre- miRNA.

- LEN attribute represents the length of the input pre-miRNA sequence.
- NBP is number of base pairs in the given sequence.
- BLR denotes the ratio of base pair to sequence length. BLR = NBP/LEN
- NHP denoted number of hairpin structure in the stem-loop structure of the given sequence.
- HPL denotes average length of hairpins in the given sequence.
- FRE is free energy in K cal/mol in folding the sequence.
- FEN is ratio of free energy to length. As length of precursors is variable in size, therefore, FEN is used for normalization purpose. FEN = FE/LEN
- AUC denotes percentage of AU contents in the sequence.

  $AUC = (n(A) + n(U))/LEN$
- MSK denotes maximum number of continuous base pairs in the given sequence. In other words it is the largest stack in the given structure.
- SDI is symmetric difference, i.e. it is difference between the lengths of both the arms of a structure.

  SD = MOD (LEN (upper arm) –LEN (lower arm))
- MBL is length of the bulge of maximum size in either arm.
- MBS is maximum bulge symmetry i.e. difference between unpaired bases in both the arms.
- MTL denotes maximum tail length. Tail is unpaired base in the start and the end of the given sequence.
- NTL is number of tails in the given sequence.

In our initial work, we have used S_A a set of 14 attributes of precursors of *Apis Mellifera* for PCA and Infogain analysis to derive dominating attributes. Similarly we have constructed decision tree based on pre-miRNA from *Bombyx moori* (Mishra and Lobiyal 2009, 2008). Further, as the work progressed and based on the encouraging results and biological significance we included two more species from hexapodes namely *Drosophila melanogaster* and *Anopheles gambiae*. After in depth analysis and literature survey we have added two more attributes to S_A to cover the features from updated datasets in miRBase repository. Therefore, S_A is extended to a set of 16 attributes. The two attributes are: SBR and MBA.

- SBR is stack base ratio. It can be derived by MSK/NBP

- MBA is maximum number of bulges in any arm.

Mature miRNA attributes was not considered in our earlier work and in this work we have tried to incorporate these attributes also**.** We defined an attribute set S_B of 9 elements, where each element represents some mature miRNA property.

S_B = ***{ARM, DFL, BPN, LNM, POP, GCC, MFE, DAS, DAE}***

The meaning of each attribute is given below.

- ARM attribute represents the location of miRNA in pre-miRNA. If miRNA is located in upper arm, we are taking value 1 and value 2 in lower arm. In some cases more than one miRNA occurs in a single pre-miRNA. In such cases we are setting ARM value to 3.
- DFL denotes distance from loop. It is the distance between the last nucleotide of hairpin to the first nucleotide in the mature miRNA.
- BPN denotes the number of base pairs between miRNA and miRNA*.
- LNM is the length of mature miRNA sequence.
- POP denotes the ratio of number of base pair to the length of miRNA.
- GCC denotes percentage of GC content in the mature miRNA sequence.
- MFE is minimum free energy for folding miRNA into secondary structure.
- DAS is dominating nucleotide at start (5' UTR).
- DAE is dominating nucleotide at the end (3' UTR).

Further to avoid separate analysis of pre-miRNA and mature miRNA we considered attributes from both the sets extended S_A and S_B to determine a set S_C of 20 attributes. This set is used for analysis of mature miRNA from pre-miRNA without looking for potential pre-miRNA. The set S_C is given below.

S_C = ***{LEN, NBP, BLR, NHP, HPL, FRE, FEN, AUC, MSK, SBR, SDI, MBL, MBA, MBS, MTL, NTL, ARM, DFL, DAS, DAE}***

A program was developed in C++ that reads the output of the phase 4.4.2 as input and produces the numeric values for all attributes. For example, the attribute values extracted from the secondary structure of *ame-mir-7 MI0001594* sequence are given in the order as they appear in set S_A. In Fig. 1 and 2 the secondary structure of *ame-mir-7* is given.

CGAGCGCCGUUGCAUGGAAGACUAGUGAUUUUGUUGUUCUACUUUCGAUAUAACAAGGAAUCACUAA UCAUCCUACAAAGGCGCUCG (*ame-mir-7 sequence*)

{87, 33, 0.37, 1, 7, –37.20, 0.42, 56, 17, 2, 3, 6, 0, 0} (Attribute values)

```
       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-mir7 in miRBase repository

```
((((((((.(((...(((.((.(((((((((((((((((.......)).))))))..))))))))).)).)))..))).))))))))
```

Fig. 2. Secondary structure of ame-mir7 using RNAfold

bmo-bantam miRNA sequence, secondary structure for its precursor and mature miRNA, and attribute values are as follows.

AGGAACUACGAAACUGGUUUUCAUAAUGAUUUGAC
AGAUUGUUUUGUAUUCUGAGAUCAUUGUGAAAG
CUAAUUUUGUUCCUGGUA

(*bmo-bantum* sequence)

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

(Secondary structure of precursor)

ugagaucauugugaaagcuaauu (Mature miRNA)

..........((....))..... Secondary structure of mature miRNA using RNAfold

Here the brackets represent base pairing and dots denote mismatches (unpaired nucleotides) in the sequence

The attribute values for $S_{(\text{bmo-bantum})}$ is as follows.

{2, 5, 22, 23, 0.957, 30.435, -0.9, 0, 4}

In similar manner attribute value of negative data sets for 20 attributes of *Drosophila melanogaster* is calculated as:

{70, 19, 0.27, 1, 12, -11.9, -0.17, 57.14, 8, 0.42, 0, 3, 5, 0, 3, 2, 2, 11, 4, 2}

Attribute Reduction and Relevance Analysis

We have used Weka software (version 3.5.8) (Fra *et al*. 2004) for Principal Component Analysis of the dataset for dimension reduction. 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). PCA transformed *Apis mellifera* dataset R : 54 × 14 into $R1$: 7 × 14 of eigenvectors (Principal Components) of dimension 14. Based on the amount of variation in dataset top five eigenvectors are selected. Further, for these five eigenvectors we did ranking of attributes based on the following formula:

$$R_{i,j} = \frac{a_{i,j}}{\sigma_j^2} \tag{5}$$

Where $R_{i,j}$ is rank of i^{th} attribute of j^{th} eigen vector. $a_{i,j}$ is i^{th} attribute of j^{th} eigen vector and σ_j^2 is variance of j^{th} eigen vector. Finally we select attributes with $R_{i,j}$ more than 1.0.

After applying PCA on the dataset attributes corresponding to eigenvectors are ranked in the order of amount of variation in dataset. We selected top five attributes with variance more than 0.1. The two attributes dropped are considered as noise since their variance is very low. The selected attribute combinations correspond to first five eigen vectors. Finally attribute weights were divided by the variance of five eigen vectors. Based on the calculated values the attributes in eigen vectors are ranked. We selected four attributes (BLR, HPL, AUC and FEN) as dominating attributes.

Weka is also used for calculating information gain of attributes in the dataset for relevance analysis. We have applied best first search method in forward direction to determine the dominating attributes. The relevant attributes derived are FEN, NBP, AUC and HPL from known dataset of *Apis mellifera* pre-miRNAs.

Further dominating attributes of mature miRNA are determined using chi square attribute evaluation, SVM attribute evaluation, infogain and consistency subset evaluation methods. Chi square attribute evaluation, SVM attribute evaluation and infogain methods use ranked based approach for searching while consistency subset evaluation method uses greedy step wise searching strategy.

Relevant attribute analysis was done for known mature miRNA sequences of *Bombyx mori*. In this we obtained relevant attributes namely DFL and GCC using chi square attribute evaluation, infogain and consistency subset evaluation methods. SVM attribute evaluation gives DFL, POP and GCC as relevant attributes. The relevant attribute analysis was also performed for all four species on 20 attributes set. The results for which are given in the following Table 1.

miRNA Classification

miRNA classification also include the first three steps used in previous section i.e. data collection, secondary structure prediction and attribute measurement. After completion of these steps, it performs the following two steps.

Table 1. Selected relevant attributes on 20 attribute set

Search criteria	Evaluation method	*Anopheles gambiae*	*Apis mellifera*	*Bombyx mori*	*Drosophila melanogaster*
Best First	Cfs subset	NHP, FEN, MTL, DFL	NBP, NHP, SBR, MBL, DFL	NHP, FEN, AUC, SBR, NTL	NHP, FEN, SBR, DFL
Ranker	Chi squared	DFL, MTL, NHP, BLR, FEN	DFL, BLR, NBP, FRE, FEN	DFL, AUC, NTL, NHP, FEN	DFL, BLR, NHP, FEN, SBR
Greedy stepwise	Consistency subset	FRE, MTL, DFL	DFL	NHP, AUC, NTL, DFL	DFL
Ranker	SVM Attribute	DFL, FEN, AUC, SBR, NHP	DFL, BLR, SBR, FEN, AUC	AUC, NHP, NTL, DFL, MBL	DFL, BLR, MSK, FRE, AUC
Ranker	Infogain	DFL, NHP, MTL, BLR, NTL	DFL, BLR, NBP, FRE, NHP	DFL, AUC, NTL, NHP, FEN	DFL, BLR, NHP, FEN, SBR

- Classification based on decision tree
- Performance evaluation

Classification based on Decision Tree

miRNA classification was done based on dominating attributes obtained from information gain analysis. We have applied a tree-based classifier J48 (a weka implementation of C 4.5 algorithm) for pre-miRNA and mature miRNA classification (Quinlan 1996). Decision trees were constructed for attribute sets S_A, S_B and S_C of all 4 species. Decision trees constructed for *Apis mellifera* (14 attribute set S_A), *Bombyx mori* (9 attribute set S_B) and *Anopheles gambiae* (20 attribute set S_C) are given in Figs. 3, 4 and 5.

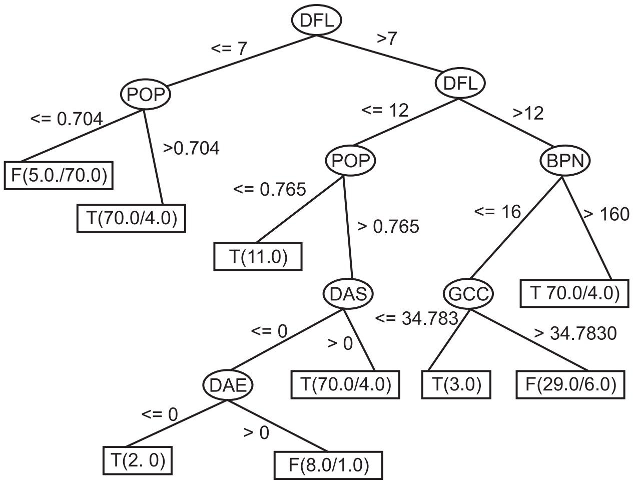

Fig. 4. Decision tree for *Bombyx mori* (9 attribute set S_B)

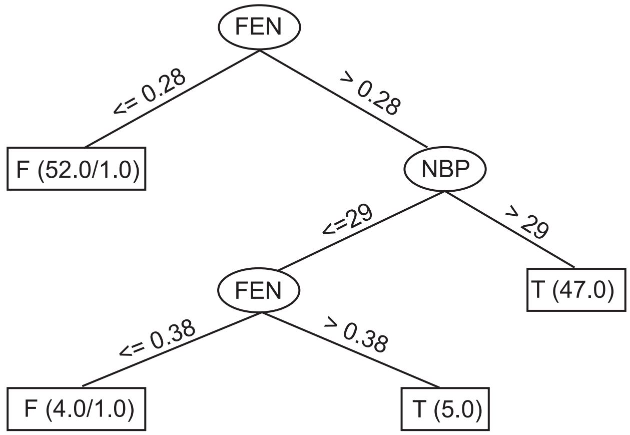

Fig. 3. Decision tree for *Apis mellifera* (14 attribute set S_A)

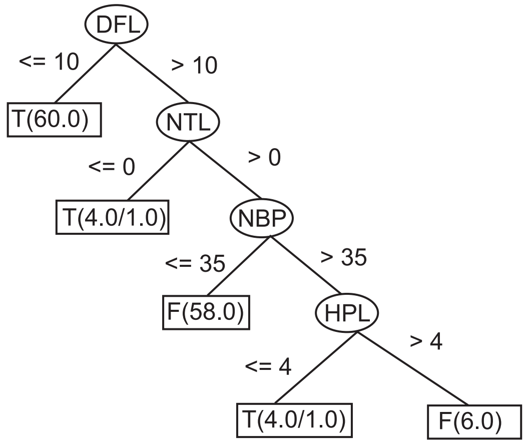

Fig. 5. Decision tree for *Anopheles gambiae* (20 attribute set S_C)

5. PERFORMANCE EVALUATION

To evaluate the performance of a model some relevant measures are used. In our work, we have used three performance measures explained below.

- Precision
- Recall
- F-measure

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), are the four different possible outcomes of a single prediction of classes True (T) and False (F). An outcome is FP when a observed class is incorrectly classified as T. When a T observed class is incorrectly classified as F it gives a FN outcome. TP and TN are correct classifications, of true and false classes, respectively.

Keeping track of all these possible outcomes (TP, TN, FP and FN) is an error-prone activity. Therefore, a confusion matrix is used to overcome the errors. All decision tree based classification use confusion matrices. In Fig. 4 confusion matrices with two classes T and F is given.

Confusion Matrix		Actual Class	
		T	**F**
Observed Class	**T**	**TP** (True Positive)	**FP** (False Positive)
	F	**FN** (False Negative)	**TN** (True Negative)

Fig. 6. Confusion matrix

Precision is statistical measure of the performance of a classification test. It is specificity of any classification model which measures the exactness of the model. Precision can be calculated as

$$\text{Precision} = TP/(TP + FP) \quad (6)$$

Recall is statistical measure of the performance of a classification test. It is sensitivity of any classification model which measures the completeness of the model. Recall can be calculated as

$$\text{Recall} = TP/\ (TP + FN) \quad (7)$$

F-measure combines precision and recall and it is the harmonic mean of precision and recall.

$$\text{F-measure} = 2*\text{Precision}*\text{Recall}/(\text{Precision} + \text{Recall}) \quad (8)$$

Standard method of evaluating the accuracy of a learning technique given a fixed sample of data is cross-validation (CV). In CV, data set is partitioned randomly into a fixed number of folds, say *n*. Each fold in turn is used for testing while the remaining ($n - 1$) folds are used for training. This procedure is repeated n times so that at the end every instance has been used exactly once for testing. This is called *n*-fold CV. To avoid random errors occurring due to data splitting, the procedure is repeated m times by choosing a different random number each time for splitting. It has been observed that many authors recommend 10 × 10 CV as a method for the best estimation of performance parameters. Therefore, 10 × 10 CV was used for experiments. Similarly in percentage split data is divided into two parts. One part is used to train the classifier and the other for the testing. In our experiment we have used percentage split to 66% means approximately 2/3 of dataset is used for training and 1/3 for testing.

We have trained the classifier with positive and negative dataset of each species with cross validation (fold 10) and percentage split (66%). After training the classifier with the dataset of a species, testing of the model was done for the data set of remaining three species. We have measured precision, recall and F-measure for 3 attribute sets of all 4 species.

For Classification using a training dataset of positive and negative samples each of 91 sequences from *Bombyx mori*, we run the model for a test set of sample size 124 from *Apis mellifera*. The result of the experiments is given in Table 2**.**

In the similar manner the classifier was also trained for positive and negative data sample each of 54 sequences from *Apis mellifera.* Testing was done on test sets of 16 and 150 sequences selected randomly from related species. The results obtained from the experiment are given in Table 3.

Table 2. Result of classification based on 9 attribute set of *Bombyx mori*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Full training set	T	0.901	0.044	0.953	0.901	0.927
	F	0.956	0.099	0.906	0.956	0.930
66 % split on training set	T	0.808	0.172	0.808	0.808	0.808
	F	0.828	0.192	0.828	0.828	0.828
CV(Fold 10)	T	0.802	0.088	0.901	0.802	0.849
	F	0.912	0.198	0.822	0.912	0.865
Test dataset	T	0.871	0.524	0.628	0.871	0.730
	F	0.475	0.129	0.784	0.575	0.692

Table 3. Result of classification based on 14 attributes set of *Apis melifera*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Full training set	T F	0.963 1.000	0 0.37	1.000 0.964	0.963 1.000	0.981 0.982
66 % split on training set	T F	0.923 1.000	0.000 0.077	1.000 0.950	0.923 1.000	0.960 0.974
Cross validation of fold 10	T F	0.944 0.044	0.056 0.056	0.944 0.944	0.944 0.944	0.944 0.944
Large test set	T F	0.964 1.000	0.000 0.036	1.000 0.957	0.964 1.000	0.982 0.978
Small test set	T F	1.000 1.000	0.000 0.000	1.000 1.000	1.000 1.000	1.000 1.000

Finally we have done classification for the set S_C (20 attribute). We trained the classifier with a species and tested it for remaining three species. The results

Table 4. Result of classification based on 20 attribute set of *Anopheles gambiae*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Training set	T F	1.000 0.970	0.030 0.000	0.971 1.000	1.000 0.970	0.985 0.985
CV with fold 10	T F	0.924 0.924	0.076 0.076	0.924 0.924	0.924 0.924	0.924 0.924
Percentage split (66%)	T F	0.870 1.000	0.000 0.130	1.000 0.880	0.870 1.000	0.930 0.936
Test against *Apis mellifera*	T F	0.968 0.984	0.016 0.032	0.984 0.880	0.968 0.984	0.976 0.976
Test against *Bombyx mori*	T F	0.879 0.791	0.209 0.121	0.808 0.867	0.879 0.791	0.842 0.828
Test against *Drosophila*	T F	0.943 0.924	0.076 0.057	0.925 0.942	0.943 0.924	0.934 0.932

Table 5. Result of classification based on 20 attributes set of *Apis mellifera*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Training set	T F	1.000 1.000	0.000 0.000	1.000 1.000	1.000 1.000	1.000 1.000
CV with fold 10	T F	0.952 0.952	0.048 0.048	0.952 0.952	0.952 0.952	0.952 0.952
Percentage split (66%)	T F	0.950 1.000	0.000 0.05	1.000 0.957	0.95 1.000	0.974 0.976
Test against *Ano.gambiae*	T F	0.894 0.924	0.076 0.106	0.922 0.897	0.894 0.924	0.908 0.910
Test against *Bombyx mori*	T F	0.868 0.813	0.187 0.132	0.823 0.860	0.868 0.813	0.845 0.836
Test against *Drosophila*	T F	0.930 0.943	0.057 0.070	0.942 0.931	0.930 0.943	0.936 0.937

obtained from each experiment are summarized in the following four Tables 4 to 7.

Table 6. Result of classification based on 20 attributes set of *Bombyx mori*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Training set	T	1.000	0.011	0.989	1.000	0.995
	F	0.989	0.000	1.000	0.989	0.994
CV with fold 10	T	0.989	0.055	0.947	0.989	0.968
	F	0.945	0.011	0.989	0.945	0.966
Percentage split (66%)	T	0.931	0.061	0.931	0.931	0.931
	F	0.939	0.069	0.939	0.939	0.939
Test against *Ano. Gambiae*	T	0.909	0.045	0.952	0.909	0.930
	F	0.955	0.091	0.913	0.955	0.933
Test against *Apis mellifera*	T	0.952	0.032	0.967	0.952	0.959
	F	0.968	0.048	0.952	0.968	0.960
Test against *Drosophila*	T	0.917	0.019	0.98	0.917	0.947
	F	0.981	0.083	0.922	0.981	0.951

Table 7. Result of classification based on 20 attributes set of *Drosophila melanogaster*

Dataset	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Training set	T	0.975	0.000	1.000	0.975	0.987
	F	1.000	0.025	0.975	1.000	0.987
CV with fold 10	T	0.968	0.000	1.000	0.968	0.984
	F	1.000	0.032	0.969	1.000	0.984
Percentage split (66%)	T	0.982	0.000	1.000	0.982	0.991
	F	1.000	0.018	0.980	1.000	0.990
Test against *Anopheles gambiae*	T	0.955	0.000	1.000	0.955	0.977
	F	1.000	0.045	0.957	1.000	0.978
Test against *Apis mellifera*	T	0.968	0.016	0.984	0.968	0.976
	F	0.984	0.032	0.968	0.984	0.976
Test against *Bombyx mori*	T	0.923	0.121	0.884	0.923	0.903
	F	0.879	0.077	0.92	0.879	0.899

6. RESULTS AND DISCUSSION

We have derived pre-miRNA attributes from their secondary structure and used different relevance analysis techniques for selecting dominating attributes. The results are encouraging since the dominating attributes selected are biologically significant. It is evident from the selected attributes-number of base pairs (NBP), number of hairpins (NHP) and free energy per nucleotide (FEN). Most stable RNA structure has larger number of base pairs with minimum free energy and a stem loop structure with a singe hairpin. In case of mature miRNA the results are also encouraging since the essential attributes selected here are also biologically significant. The attributes selected are DFL, POP and GCC represents distance from loop, percentage of pairing and GC content respectively. Through the analysis of mature miRNA it is evident that miRNAs has high percentage of base pairing and GC rich regions mostly located near hairpins.

The precision and recall measures on training and test datasets are quite satisfactory. The model has worked well on dataset of the same species. The results obtained from experiments clearly show that our model gives high precision and recall in all cases. However, the model is not able to recall few true cases, which may be improved by taking a larger training dataset. The large training dataset may also help in identifying dominating attributes for increasing the accuracy of prediction model.

7. CONCLUSION

The results of attribute relevance analysis are encouraging since the essential attributes selected using different techniques are biologically significant. The high values obtained for precision and recall validate accuracy of the classification model as well as relevance of dominating attributes.

However, the results obtained here can be further, validated by using other methods of miRNA prediction. Similarly the significance of dominating attributes can also be verified by using other techniques of attribute reduction and selection. In addition to decision tree based classification, other techniques of classification may also be applied to improve the accuracy and efficiency of classification.

REFERENCES

Bartel, D.P. (2004). MicroRNAs: Genomics, biogenesis, mechanism and function. *Cell,* **116**, 281-297.

Cullen, B.R. (2004). Transcription and processing of human microRNA precursors. *Mol. Cell*, **16**, 861-865.

Fra, E., Hall, M. *et al*. (2004). Mining in bioinformatics using Weka. *Bioinformatics,* **20(15)**, 2479-2481.

Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, The University of Waikato.

Kohavi, R. (1995). Wrappers for Performance Enhancement and Oblivious Decision Graphs. PhD thesis, Stanford University.

Kohavi, R. and John, G. (1996). Wrappers for feature subset selection. A*rtificial Intelligence, Special Issue on Relevance*, **97**, 273-324.

Lee, R.C., Feinbaum, R.L. *et al.* (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* **75**, 843-854.

Lee, Y., Ahn, C., Han, J. *et al.* (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415-419.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W. *et al.* (2003). Prediction of mammalian microRNA targets. *Cell,* **115**, 787-798.

Mathews, D.H., Sabina, J., Zuker, M. *et al.* (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.,* **288**, 911-940.

Mishra, A.K. and Lobiyal, D.K. (2008). Exploring dominating features from Apis Mellifera pre-miRNA. International Conference on Advanced Computer Theory and Engineering *(ICACTE-2008 Phuket, Thailand)*, http://doi.ieeecomputersociety.org/10.1109/ICACTE. **169**, 363-367.

Mishra, A.K. and Lobiyal, D.K. (2009). Apis mellifera pre-miRNA prediction using decision tree based classifier. International Conference on Computer and Automation Engineering *(ICCAE-2009, Bangkok, Thailand)* ISBN: 978-0-7695-3569-2, http://doi.ieeecomputersociety.org/10.1109/ICCAE., **5**, 123-126.

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.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, **1**, 81-106.

Quinlan, J.R. (1996). Improved use of continuous attributes in C4.5. *J. Artificial Intell. Res.*, **4**, 77-90.

Quinlan, J.R. (1987) Simplifying decision trees. *Intern. J. Man-Machine Studies*, **27**, 221-234.

Quinlan, J.R. (1989). Inferring decision trees using the minimum description length principle. *Infor. Comput.*, **80**, 227-248.

Tinoco Jr., I., Uhlenbeck *et al.* (1971). Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362-367.

Witten, I.H. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., **9**, 133-148.