Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 3, Issue 1, 2012, pp 281-289 http://www.bipublication.com

MACHINE LEARNING MODEL FOR CLASSIFICATION OF MOTIFS IN HUMULUS (HOP)

Swati Singh[1*], Ashutosh Mani[1*], Anubha Dubey[2], Anoop Chaturvedi[3]
[1]Center of Bioinformatics, University of Allahabad, Allahabad, India-211002;
[2]Department of Bioinformatics, MANIT Bhopal;
[3]Department of Statistics, University of Allahabad, Allahabad, India-211002.
*corresponding author:- Email-swatinatural@gmail.com;

ABSTRACT

Protein sequence motifs are signatures of protein families and can often be used as tools for the prediction of protein function. There is need for development of computational technique for prediction and classification of motifs in Humulus Lupulus and Humulus japonicus protein sequences. Here Machine Learning algorithms are used in developing model for classification of motifs of genus Humulus from four different Profiles or database. Various algorithms have been used to classify motifs in sequences but the fair accuracy is observed from J48, Bayesnet, and RandomForest. The Information extracted from these models can be of great use in predicting functions and evolution.

Keywords – Humulus, classification, motifs, J48, BayesNet, RandomForest

INTRODUCTION

Humulus is a genus of dioecious, anemophilous, dextrose – twining herbaceous vines, indigenous in north temperate areas. It is found in river and stream bottoms, thickets, hedgerows, and roadsides. The hop is a part of family Cannabaceae. The female flowers (often called 'cones') of the species are knows as hops and are used as culinary flavouring and stabilizer, especially in the brewing of beer.Dried female strobiles of hops have historically been used for its sedative effects on the central nervous system due to its methylbutenol content as a treatment for stress, anxiety and insomnia, they have anti-viral properties and anti-micro bacterial properties, they have been studied for containg estrogen precursors as well. Three different species of the genus *Humulus* [1] are recognized. Each differ slightly to each other by the range of habitat it can live in, growth rate, and the amount of chemicals in the strobiles. *Humulus lupulus, Humulus japnicus, Humulus Yunnanesis.*

Protein sequence classification constitutes an important problem in biological sciences for annotating new protein sequences and detecting close evolutionary relationships among sequences. Here we have developed machine learning models in classifying motifs from four Profiles in *Humulus Lupulus* and *Humulus Japonicus*. However there is no sequencing done for *Humulus Yunnanesis*. Motifs [2] are the recurring element in biosequence and structure and a motif typically has functional implications since it is preserved (recurring) during evolution. There are different types of motifs-sequence motif, structure motifs and network motifs. Protein sequence motifs are generally functional sites. So we have taken motif based protein sequence classification [2, 3].

There are a number of databases for retrieving and predicting functional motifs in protein sequences. There are various software for motif identification such as InterProScan, MyHits, ScanProsite, HamapScan, MotifScan, PPSEARCH, SMART, PATTINPROT, PRINTS, ProDom [5]. We used four database/profile for searching of motifs- Prosite Profile [6], HAMAP Profile [7] and PfamHMM local model and PfamHMM global model [8] and different parameters (attributes) of motifs are used from these four profiles, Sart region, End region, N-Score E-Score and Species.

Five Parameters of motifs are as follows:-

1. *Start*-Starting position of motif
2. *End*-End position of motif

3. E-Value-The E-value is the number of matches with a score equal to or greater than the observed score that are expected to occur by chance. In other words, the E-value provides an estimation of the number of false positives. The E-value depends on the size of the database searched, as the number of false positives expected to be above a given score threshold usually increases proportionately with the size of the database.

4.N-Score **-** The normalized score are defined as the base 10 logarithm of the size (in residues) of the database in which one false positive match is expected to occur by chance.N-Score is independent of the size of the databases For a given database size of DB_size residues, the normalized score N_score and the E-value are easily interconvertible:

N_Score = log10 DB_size -log10 E-value

or

E-value = DB_size*10-N_Score

5. *Species* – Two species are taken in input *Humulus Lupulus* as plant1 and *Humulus Japonicus* plant 2

In this paper various machine learning agorithms such as BayesNet, RandomForest, J48, Logistic, IB1, have been used in classification of functional sites (motifs) in HAMAP Profile, Prosite profile, PFAM HMM local model and PFAM HMM global Model of *Humulus Lupulus* and *Humulus japonicus.*

METHODLGY

Here the protein sequences of genus *Humulus* are extracted from UniProtKB/Swiss-Prot database [9] from which the classification of motifs-based on four profiles is the main objective of this paper. There are various machine learning algorithms [10] are available for the classification and prediction of motifs from different profile. Model has been developed using different algorithms of WEKA classifier [11]. They give different result, accuracy and time for model preparation. This variation in result and accuracy leads to dilemma of choosing algorithm for classification and prediction of motifs. Classification [12, 13] is a basic task in data analysis and pattern recognition that requires the construction of a classifier a function that assigns a class label to instances described by a set of *attributes*. From the various algorithms BayesNet, J48, Random Forest gives the better result with fair accuracies.

BayesNet - Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute A given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A_i …..A_n and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes [14].

J48- J48 Decision tree classifier follows simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals [15].

RandomForest – It is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [16].

RESULTS:

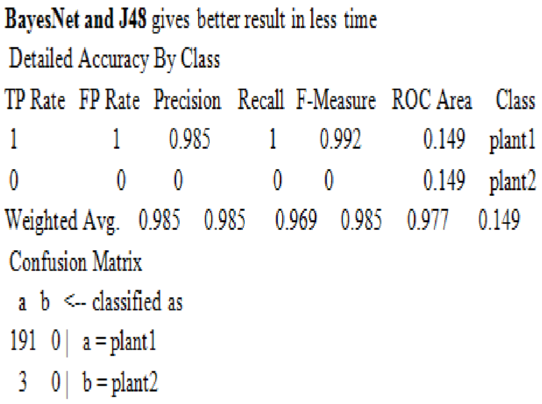
For developing model for Classification we obtained our dataset from UniProtKB /Swissprot database and from four different profile of motif finding HAMAP, PROSITE, PfamHMM local and Pfam HMM global obtained motifs from both species *Humulus Lupulus* and *Humulus Japonicus*. Classification of motifs of genus *Humulus* on the four profiles of motif finding by BayesNet, J48, RandomForest will give the following results with fair accuracy and with minimum time in model preparation.

Class I. Result for Classification of motifs from four profile (Start,End,N-Score and E-value)

a) **HAMAP Profile**

Table1

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0.02
J48	98.4536	0.985	0.02
RandomForest	98.4536	0.985	003



```
BayesNet and J48 gives better result in less time
 Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
 1         1       0.985      1      0.992       0.149    plant1
 0         0       0          0      0           0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.149
 Confusion Matrix
  a   b   <-- classified as
 191  0 |  a = plant1
  3   0 |  b = plant2
```

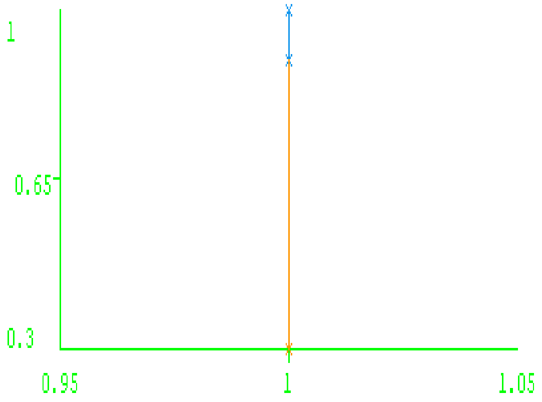


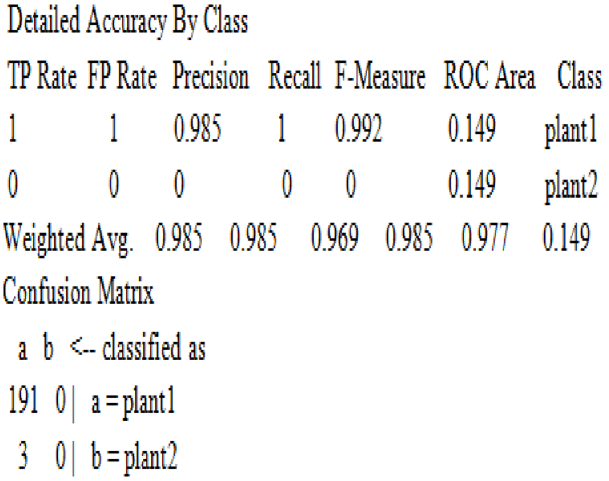
Figure1 : ROC of all parameters of motifs

b)PfamHMM global model

Table 2

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0
J48	98.4536	0.985	0.02
RandomForest	97.9381	0.979	0.02

BayesNet gives fair accuracy in less time



```
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
1         1        0.985      1       0.992      0.149     plant1
0         0        0          0       0          0.149     plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.149
Confusion Matrix
 a  b  <-- classified as
191  0 |  a = plant1
 3   0 |  b = plant2
```

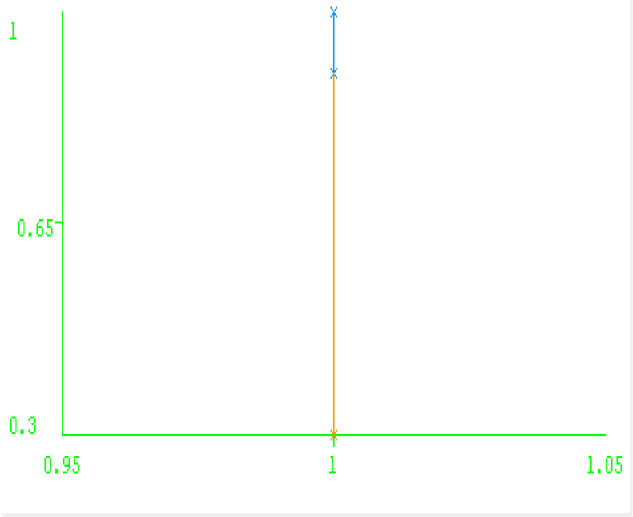
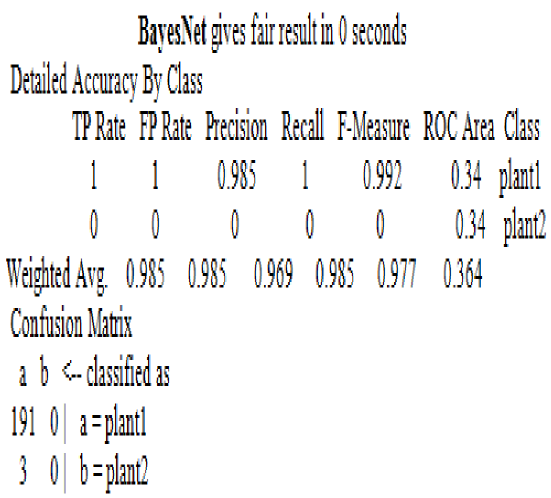


Figure2 : ROC of all parameters of motifs

c) PfamHMM local model

Table 3

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0
J48	98.4536	0.985	0.02
RandomForest	98.4536	0.985	0.02



```
              BayesNet gives fair result in 0 seconds
Detailed Accuracy By Class
          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
            1        1        0.985      1       0.992      0.34    plant1
            0        0        0          0       0          0.34    plant2
Weighted Avg.  0.985  0.985   0.969    0.985   0.977      0.364
Confusion Matrix
 a  b  <-- classified as
191  0 |  a = plant1
 3   0 |  b = plant2
```

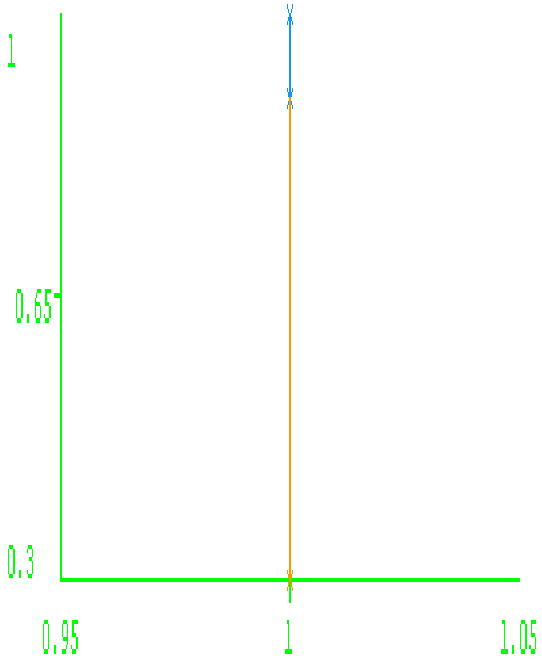
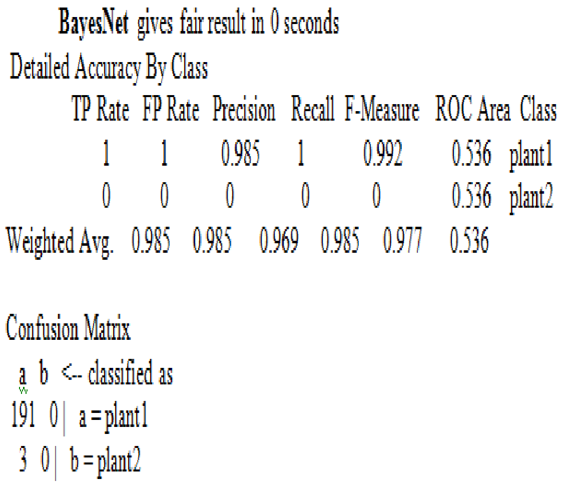


Figure3 : ROC of all parameters of motifs

d)Prosite Profile
Table 4

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0.02
J48	98.4536	0.985	0.02
RandomForest	98.4536	0.985	0.03



```
              BayesNet  gives fair result in 0 seconds
Detailed Accuracy By Class
          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
            1        1        0.985      1       0.992      0.536   plant1
            0        0        0          0       0          0.536   plant2
Weighted Avg.  0.985  0.985   0.969    0.985   0.977      0.536

Confusion Matrix
 a  b  <-- classified as
191  0 |  a = plant1
 3   0 |  b = plant2
```

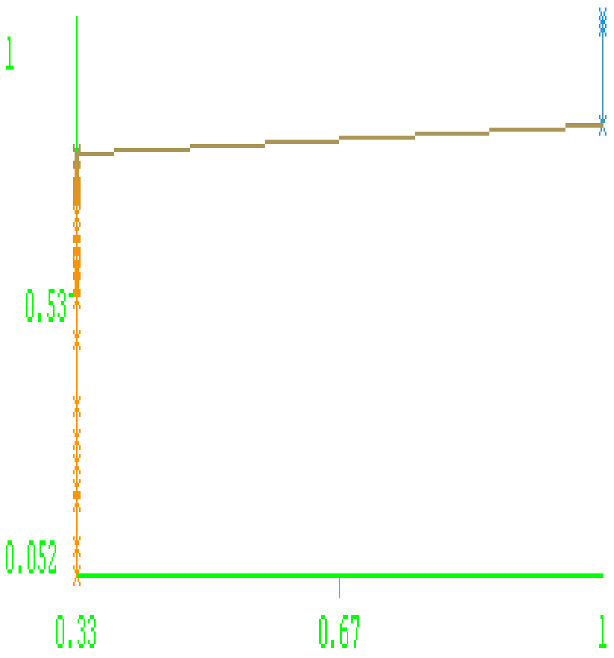


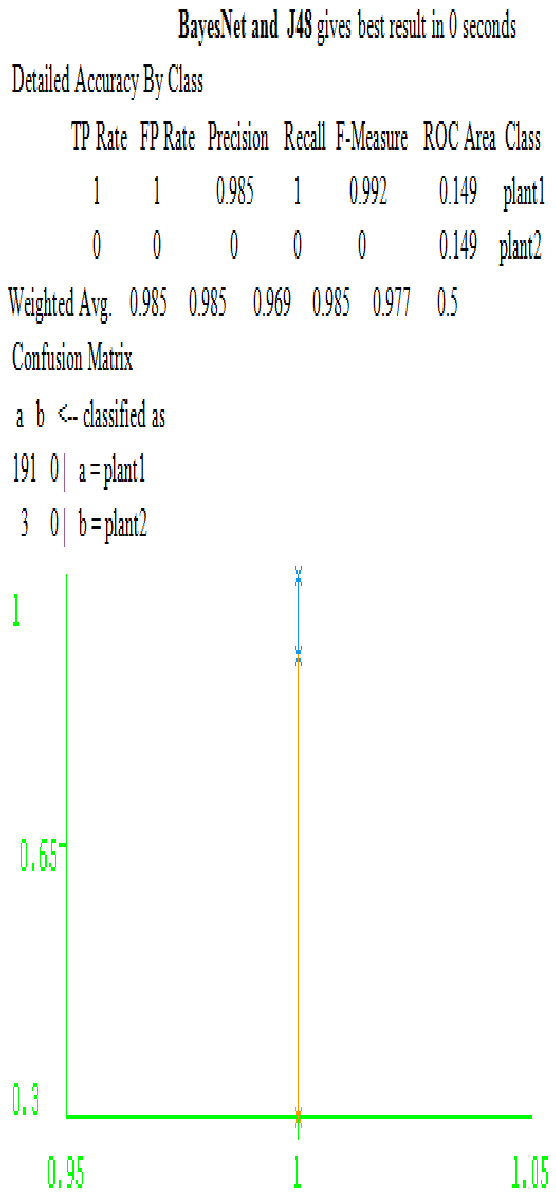
Figure4 : ROC of all parameters of motifs

Class II. Result for Classification of motifs from four profile (Start,End and E-value)

a)**HAMAP Profile**

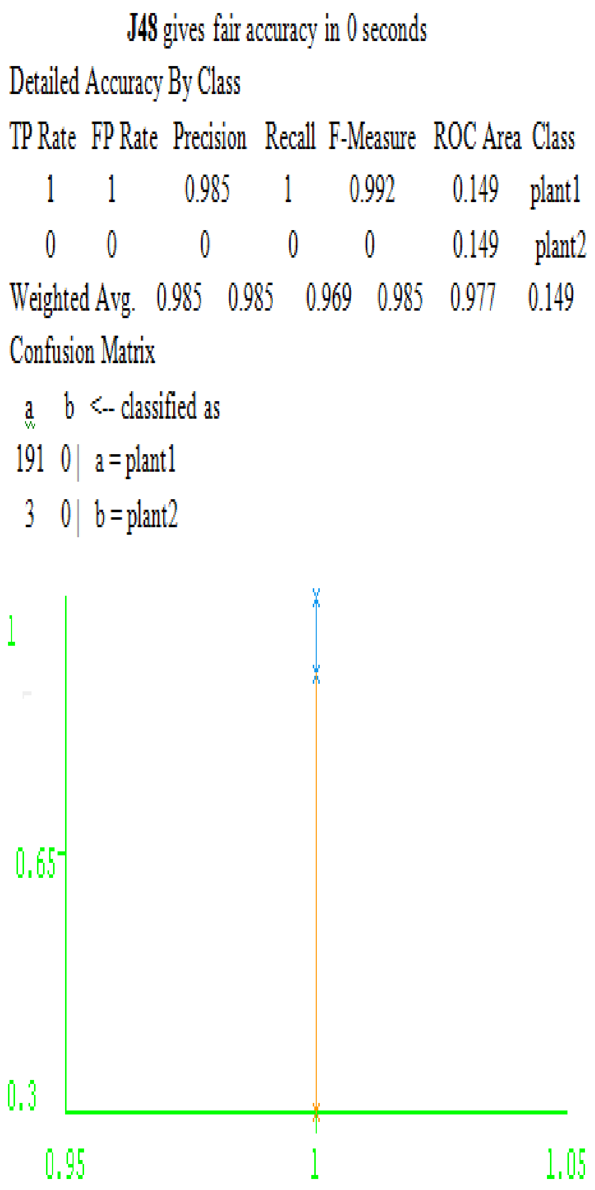
Table 5

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0
J48	98.4536	0.985	0
RandomForest	98.4536	0.985	0.02



```
            BayesNet and J48 gives best result in 0 seconds
Detailed Accuracy By Class
        TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          1        1        0.985      1       0.992      0.149    plant1
          0        0        0          0       0          0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.5
Confusion Matrix
 a  b  <-- classified as
191  0|  a = plant1
 3   0|  b = plant2
```

Figure5: ROC of 3 parameters of motifs (Start,End and E-Score)



```
               J48 gives fair accuracy in 0 seconds
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
   1        1       0.985       1      0.992      0.149    plant1
   0        0       0           0      0          0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.149
Confusion Matrix
 a   b  <-- classified as
191  0|  a = plant1
 3   0|  b = plant2
```

Figure6: ROC of 3 parameters of motifs (Start,End and E-Score)

b)PfamHMM global model

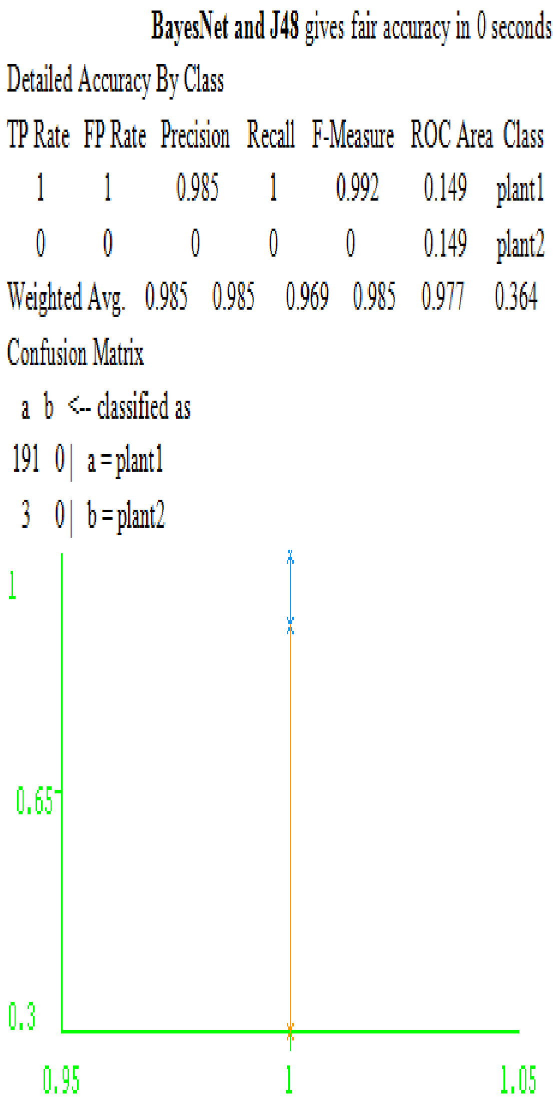
Table 6

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
J48	98.4536	0.985	0
BayesNet	98.4536	0.985	0.02
RandomForest	97.9381	0.979	0.02

c) Pfam-HMMlocal model

Table 7

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
J48	98.4536	0.985	0
BayesNet	98.4536	0.985	0
RandomForest	97.9381	0.979	0.02



BayesNet and J48 gives fair accuracy in 0 seconds

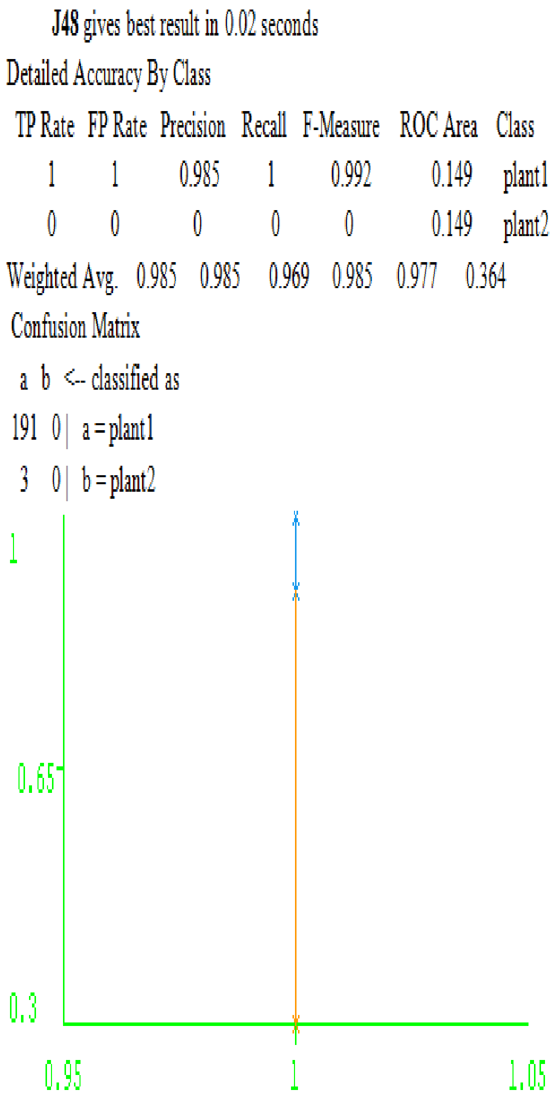
```
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
   1        1       0.985       1      0.992     0.149    plant1
   0        0       0           0      0         0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.364
Confusion Matrix
  a   b  <-- classified as
 191  0 |  a = plant1
  3   0 |  b = plant2
```

Figure7: ROC of 3 parameters of motifs (Start,End and E-Score)

d)Prosite Profile

Table 8

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0.02
J48	98.4536	0.985	0
RandomForest	98.4536	0.985	0.02



J48 gives best result in 0.02 seconds

```
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
   1        1       0.985       1      0.992     0.149    plant1
   0        0       0           0      0         0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.364
Confusion Matrix
  a   b  <-- classified as
 191  0 |  a = plant1
  3   0 |  b = plant2
```

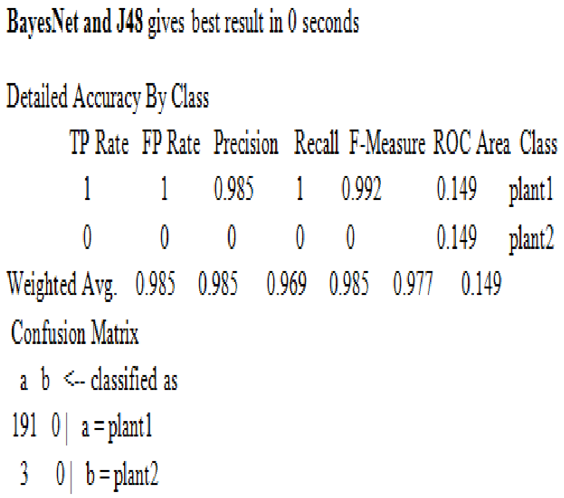
Figure8: ROC of 3 parameters of motifs (Start,End and E-Score)

Class III. Result for Classification of motifs from four profile (Start,End and N-score)

a)HAMAP Profile

Table 9

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0
J48	98.4536	0.985	0
DecisionTable	98.4536	0.985	0.02



```
BayesNet and J48 gives best result in 0 seconds

Detailed Accuracy By Class
          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
           1        1        0.985      1       0.992      0.149    plant1
           0        0        0          0       0          0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.149
Confusion Matrix
 a   b   <-- classified as
191  0 |  a = plant1
 3   0 |  b = plant2
```

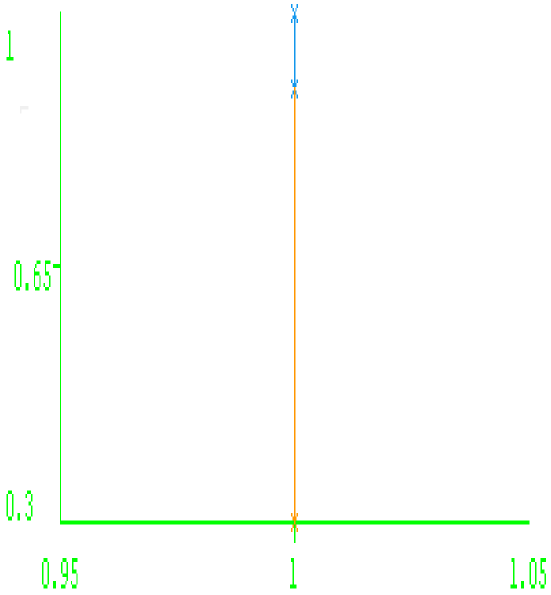
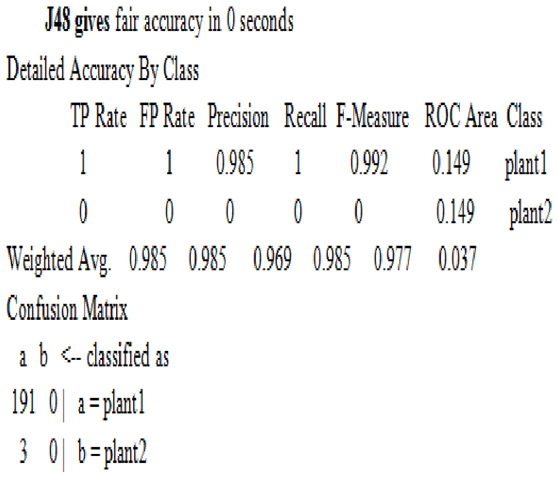


Figure9: ROC of 3 parameters of motifs (Start,End and N-Score)

b) PfamHMM global model

Table 10

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0.02
J48	98.4536	0.985	0
RandomForest	98.4536	0.985	0.02



```
J48 gives fair accuracy in 0 seconds
Detailed Accuracy By Class
          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
           1        1        0.985      1       0.992      0.149    plant1
           0        0        0          0       0          0.149    plant2
Weighted Avg.  0.985  0.985  0.969  0.985  0.977  0.037
Confusion Matrix
 a   b   <-- classified as
191  0 |  a = plant1
 3   0 |  b = plant2
```

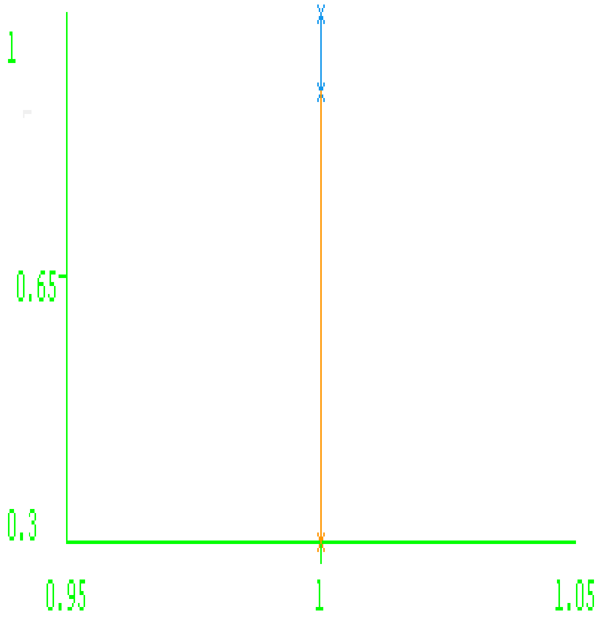
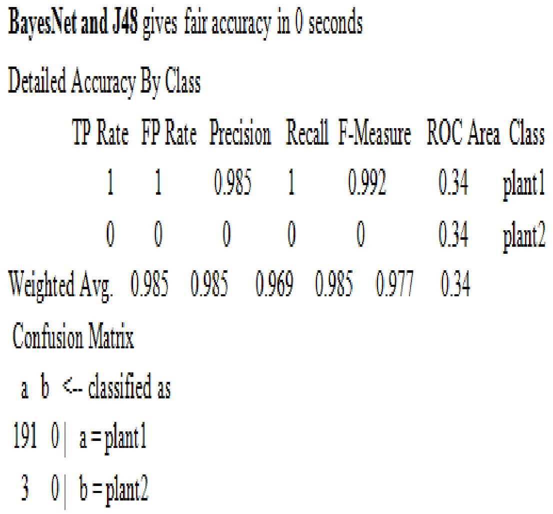


Figure10: ROC of 3 parameters of motifs (Start,End and N-Score)

c) Pfam HMM local model

Table 11

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
J48	98.4536	0.985	0
BayesNet	98.4536	0.985	0
RandomForest	98.4536	0.985	0.02



BayesNet and J48 gives fair accuracy in 0 seconds

```
Detailed Accuracy By Class
           TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
              1       1       0.985      1       0.992      0.34     plant1
              0       0       0          0       0          0.34     plant2
Weighted Avg.  0.985   0.985   0.969   0.985   0.977   0.34

Confusion Matrix
  a  b  <-- classified as
191  0 |  a = plant1
  3  0 |  b = plant2
```

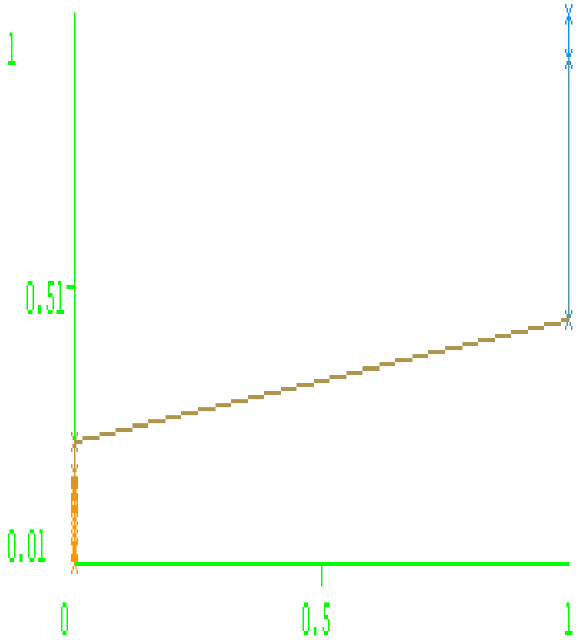
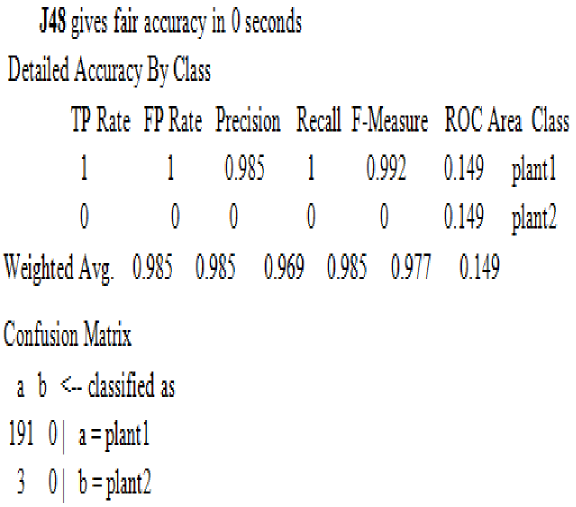


Figure11: ROC of 3 parameters of motifs (Start,End and N-Score)

d)Prosite Profile

Table 12

ALGORITHM	ACCURACY	AVERAGE	TIME TAKEN (in seconds)
BayesNet	98.4536	0.985	0.02
J48	98.4536	0.985	0
RandomForest	98.4536	0.985	0.02



J48 gives fair accuracy in 0 seconds

```
Detailed Accuracy By Class
           TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
              1       1       0.985      1       0.992      0.149    plant1
              0       0       0          0       0          0.149    plant2
Weighted Avg.  0.985   0.985   0.969   0.985   0.977   0.149

Confusion Matrix
  a  b  <-- classified as
191  0 |  a = plant1
  3  0 |  b = plant2
```

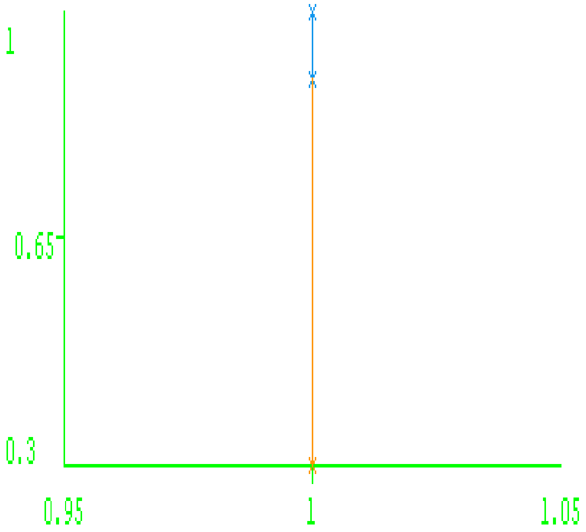


Figure12: ROC of 3 parameters of motifs (Start,End and N-Score)

DISCUSSION:

We divided the data into three classes-Class I, Class II and Class III where Class I have all parameters of motifs, Class II have three parameter excluding N-Score and Class III have three parameter excluding E-Score. Then we classified all these three classes from four profiles (HAMAP, PROSITE, PfamHMM global and PfamHMM local) by using various machine learning algorithm from which–BayesNet and J48 gives fair accuracy in HAMAP profile whereas in rest of the three profiles Bayes net shows maximum accuracy in Class I data. Both Class II and III shows same types of classification result in HAMAP profile and Pfam

HMM local model BayesNet and J48 algorithm shows maximum accuracy in minimum time and in Pfam HMM global model and Prosite J48 algorithm gives fair accuracy in classification. An ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space. Reciever Operating Curve is a graphical technique in evaluating data mining schemes. For each fold of a 10 fold cross validation, weight the instances for a selection of different cost ratios train the scheme on each weighted set, count the true positives and false positives in the test set, and plot the resulting point on the ROC axes. The ROC curves for different classes have been plotted from Figs (1-12). As ROC depicts the performance ,we can refer from the confusion matrix that in all class of motif finding best algorithm shows TPR and FPR 1. The accuracy of results for the three class having four profiles obtained from all the above mentioned classifiers with input as motif parameters predicted from different three classifier and their comparison is presented in Tables (1-12). In our classes we have choosen three classifier giving same accuracy 98.4536 and less time taken for model preparation.

CONCLUSION:

Among all the classifier, classification of motifs with Start, End, E-Value and N-Score of Class I of four profile BayesNet found suitable. In Class II and Class III we found that J48 gives suitable classification in less time in four different profiles. The accuracy of model is maintained as more proteins sequences were discovered in Humulus. As motifs are conserved in protein sequences and has functional implications, it is preserved during evolution. Therefore we can perform evolution based study in classification of proteins.

ACKNOWLEDGEMENT:

First Author is thankful to University Grants Commission, New Delhi for a research fellowship.

REFRENCES:

[1] Humulus lupulus (2003),Alternative Medicine Review Volume 8, Number 2 .

[2] Brazma, A., Jonasses, I., Eidhammer, I., andGilbert, D.(1998). Approaches to the automatic discovery of patterns in biosequences. Journal of Computational Biolog y, 5(2):277–303.

[3] Blekas K, Fotiadis DI, Likas A (2005). Motif-based protein sequence classification using neural networks. J Comput Biol. ;12(1):64-82.

[4] Sotiris Diplaris, Grigorios Tsoumakas, Pericles A. Mitkas, and Ioannis Vlahavas , (2005).Protein Classification with Multiple Algorithms P. Bozanis and E.N. Houstis (Eds.): PCI 2005, LNCS 3746, pp. 448 – 456

[5] http://expasy.org/proteomics/families patterns and profiles.

[6]Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch and Philipp Bucher (2002),” PROSITE: A documented database using patterns and profiles as motif descriptors”. BRIEFINGS IN BIOINFORMATICS. VOL 3. NO 3. 265–274.

[7]Tania Lima, Andrea H. Auchincloss, Elisabeth Coudert (2009),” HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot” Nucleic Acids Research, Vol. 37.

[8] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn The Pfam protein families database **Nucleic Acids Research** (2012) Database Issue 40:D290-D301.

[9] http://www.uniprot.org/help/uniprotkb

[11] http://www.cs.waikato.ac.nz/ml/weka

[12] Duad, R., Hart, P. (1973): Pattern Classification and Scene Analysis. Wiley, New York

[13] F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.) (2007): Biomed 06, IFMBE Proceedings 15, pp. 520-523.

[14] Remco R. Bouckaert Bayesian Network Classifers in Weka.

[15] Quinlan, R. (1993): C4.5: Programs for Machine Learning. Morgan Kaufman, San Mateo .

[16] Pang-Ning, Tan.M.Steinbach, V.Kumar (2008). Introduction to Data Mining.