I J B R

International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 2, 2011, pp-202-206
Available online at http://www.bioinfo.in/contents.php?id=21

SUPPORT VECTOR MACHINE FOR CLASSIFICATION OF HIV, PLANT AND ANIMAL miRNA'S

ANUBHA DUBEY*[1] AND USHA CHOUHAN[2]
[1]Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal, India
[2]Department of Mathematics, Maulana Azad National Institute of Technology, Bhopal, India
*Corresponding Author: Email- anubhadubey@rediffmail.com

Received: August 02, 2011; Accepted: August 28, 2011

Abstract- MicroRNAs (miRNA's) constitute a large family of non coding RNAs that function to regulate gene expression. Wet lab experiments usually used to classify the miRNA of plants and animals are highly expensive, labor intensive and time consuming. Thus there arises a need for computational approach for classification of plant and animal miRNA. These computational approaches are fast and economical as compared to wet lab techniques. Here a machine learning approach is used to classify miRNA of HIV, plants and animals. The new SVM learning algorithm called Weka LibSVM has been used for classification of plant and animal and HIVmiRNA. The model has been tested on available data and it gives results with 95% accuracy.

Keywords- MicroRNAs, Kernel, Hyper plane, Support Vector

Introduction

MicroRNAs (miRNA's) are small RNAs of 21–25 nucleotides that specifically regulate cellular gene expression at the post-transcriptional level. miRNA's are derived from the maturation by cellular RNases III of imperfect stem loop structures of ~ 70 nucleotides. Evidence for hundreds of miRNA's and their corresponding targets has been reported in the literature for plants, insects, invertebrate animals, and mammals. The analysis of miRNA-encoding potential to the human immunodeficiency virus (HIV) is being studied [1]. Using computer-directed analyses, it was found that HIV putatively encodes five candidate pre-miRNA's [Fig 1]. Folded pre-miRNA and their corresponding predicted mature viral miRNA (red) are listed. Nucleotide positions (where 1 is the initiation of transcription) in the pNL4-3 genome are presented in the right column.It was then matched deduced mature miRNA sequences from these 5 pre-miRNA against a database of 3' untranslated sequences (UTR) from the human genome. These searches revealed a large number of cellular transcripts that could potentially be targeted by these viral miRNA (vmiRNA) sequences [1].

Correct identification of miRNA that regulate cellular processes and impact economically important traits is the need of the industry. This requires better understanding of characteristics of miRNA's which can be done by understanding the differences between miRNA's of different organisms [2], [8]. They have applications 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. As of now various attempts have been made to discover novel miRNA's in various species of plants, animals and viruses 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, animal [15] and HIV miRNA's, 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.

For the classification to be successful, each class must show some distinct properties or characteristics. 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 but many dissimilarity with HIV exist listed in table 1below which are used used in the classifier.

MODEL AND METHOD

SVM are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that one maximizes maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data

sets [11, 12].Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier [8], [9], [10].A support vector machine (SVMs) is a useful technique for data classification. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one target value (class labels) and several attributes (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only
the attributes [4].

A. Software

The miRNA target registry software is used to extract the properties of the animal and plant miRNA [13]. Weka and LibSVM are two efficient software tools for building SVM classifiers. Each one of these two tools has its points of strength and weakness. Weka has a GUI and produces many useful statistics (e.g. confusion matrix, precision, recall, Fmeasure, and ROC scores) [6]. LibSVM runs much faster than Weka SMO and supports several SVM methods (e.g. one-class SVM, nu-SVM, and R-SVM). Weka LibSVM (WLSVM) combines the merits of the two tools. WLSVM can be viewed as an implementation of the LibSVM running under Weka environment [2].

B. Evaluation

There are two parameters while using RBF kernels: C and γ (radial basis function: exp (-gamma*|u-v|^2). It is not known beforehand which C and γ are the best for one problem; consequently some kind of model selection (Parameter search) must be done. The goal is to identify good values so that the classifier can accurately predict
unknown data (i.e., testing data) to achieve which, we perform a grid-search on C and gamma using 10-fold cross validation to accurately predict unknown data. Here is the result of our grid search shown in figure 2 on one of our training set to find best C and gamma using the RBF kernel function.

The value of c=2 and gamma= 0.0078125 obtained using grid search was implemented in weka Libsvm which gave 95% accurate results. In the two class case with classes yes and no, a single prediction has the four different possible outcomes (where TP = true positive, FP = false positive, TN = true negative and FN = false negative). The true positive and true negative is correct classification. The overall success rate is the number of correct classification divided by the total number of classification.

$$Accuracy(x) = \frac{tp + tn}{tp + tn + fp + fn}$$

Finally the error rate is one minus success rate.

Confusion matrix

```
A       b <-- classified as
81     1 | a = Animal
41      4 | b = Plant
```

As data of HIV miRNA is not sufficient for the software so it excludes it and classifies only plant and animal miRNA. In a multiclass prediction the result on a test set is often displayed as a two dimensional confusion matrix with a row and column for each class. In matrix element the actual class is the row and the predicted class is the column. In this case test set has 100 instances (81+14=95) of them are predict correctly so the success rate is 95%.

C. Training set

For the classification of animal, plant and HIV miRNA we select the dissimilarity between the animal and plant miRNA from table1. For the classification purpose we have selected the characteristics, cluster members in both plants and animals. As the data of HIV miRNA is very small so the software considers only plant and animal miRNA data. The cluster members can have the values common/uncommon. Clusters are generally found in animals whereas very less plant species contain them. Number of mismatches in animal and plant is the second feature used in our classifier with different numeric values. 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. The number of target genes is one more feature included in the classifier [7]. Animals generally show large number of targets belonging to different families but plants have less number of targets generally belonging to one family. Size of fold back loop is greater than 100 for plant with variations till 303 nucleotides and less than 100 nucleotides for animals [13]. With these characteristics we have trained the WEKA classifier and the values we get are given in the table 2.

RESULT AND DISCUSSION

A set of attributes was collected from miRNA registry [13] 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. First, some attributes are clearly not independent and do not provide any additional advantage when evaluated together [5]. For example complimentarily is a feature which is dependent on number of mismatches with the target. The result show 95% classified instances and only 5% unclassified instances. The detailed characteristics of the classification are given in Table II which uses support vector machine. The values we get from the curve for plant and animal are 0.222 and 0.012 which lies in the false positive region and the values of ROC curve for this data set comes out to be 0.883 for animal and plant in true positive region.

CONCLUSION

The four characteristics (number of mismatches with target mRNA, presence of clusters, number of target genes and the size of fold back loop) used in the SVM model to develop the classifier give us fairly good accuracy in results i.e. 95% accuracy. Thus we infer that these four characteristics are very important and must be included in any classifier of plant animal and HIV miRNA's. Apart from these there are characteristics which were dependent and their inclusion in the model does not bring any improvement. 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.

ACKNOWLEDGMENT

The authors are highly thankful to the Department of Biotechnology, New Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinformatics infrastructure facility to carry out the research work.

REFERENCES

[1] Yamina Bennasser, Shu-yun Le, Man Lung Yeung and Kuan-Teh Jeang (2004) *Retro virology*, 1:43doi:10.1186/1742-4690-1-43

[2] Millar A. A., Waterhouse P. M. (2005) *Springer Link Function Integral Genomics*, vol. 5, pp.129-135.

[3] Chang C.-C. and Lin C.-J. (2001) *LIBSVM: a library for Support Vector Machines, (2001).Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.*

[4] Aagaard L. and John J. Rossi (2007) *Elsevier Science, Advanced Drug Delivery Reviews* 59, 75–86.

[5] Witten I.H., Frank E. (2005) *Data Mining – Practical machine learning Tools and techniques with Java implementations, Morgan Kaufmann, San Francisco*.

[6] De Ferrari L. and Aitkin S. (2006) *BMC Genomics*, 7:277.

[7] *Weka Data Mining Java Software, http://www.cs.waikato.ac.nz/~ml/.weka/.*

[8] Jones-Rhoades M.W., Bartel D.P. (2004) *Mol Cell* 14:787–799.

[9] Langley P. and Sage S. (1994) *Elements of machine learning. And Fracisco: Morgan Kaufmann*.

[10] Han J. and Kamber M. (2004) *Data mining: concepts and Techniques.*

[11] Holte R.C. (1993) *Very simple classification rules perform well on most commonly used datasets .Machine Learning*.

[12] Cortes C., Vapnik V. (1995) 20:273-297.

[13] Burges C. J. C. (1998) *Data Mining and Knowledge Discovery, A Tutorial on Support Vector Machines for Pattern Recognition*.

[14] *Micro RNA Registry www.misangar.ac.uk*

[15] Kumud Pant, Bhasker Pant, Pardasani K.R. (2009) *International Conference on Advances in Computing, Control, and Telecommunication Technologies. Published by IEEE Computer Society in IEEE Xplore USA*, 338 – 340.

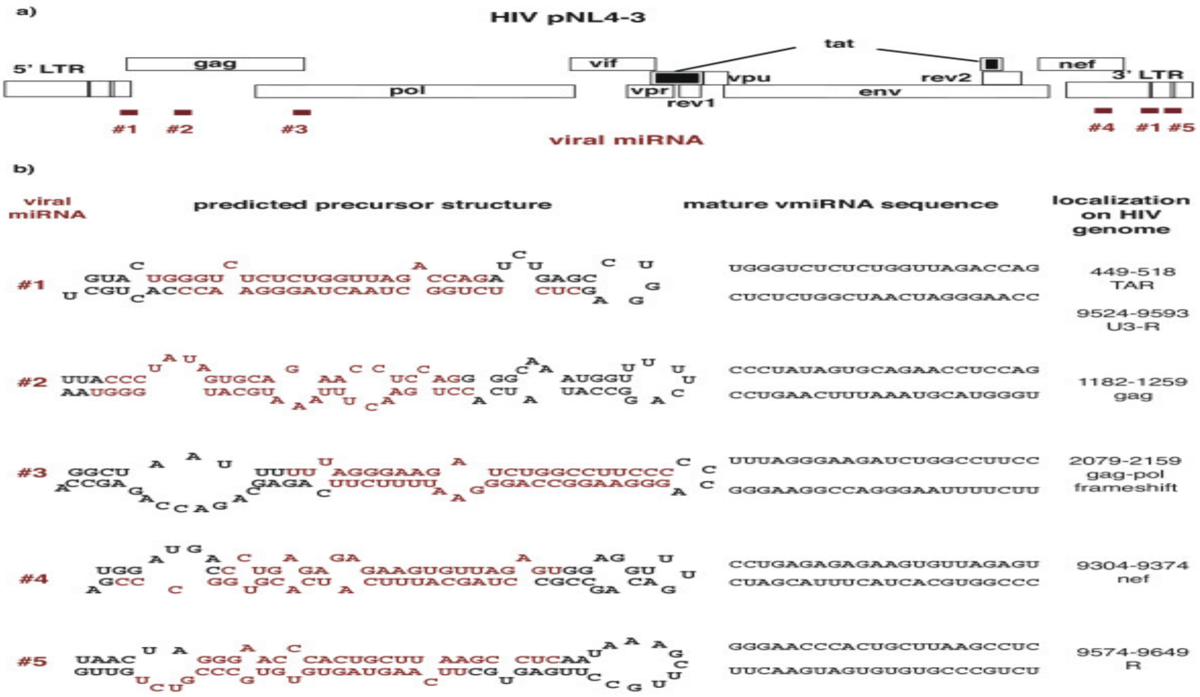


Fig. 1- Sequences and localization of HIV-encoded miRNA **candidates. a)** locations for 5 predicted pre-miRNA's candidates in the pNL4-3 genome are shown. **b**) The folded structures of the 5 viral pre-miRNA from pNL4-3 (Accession Number AF 324493) [16] are illustrated.

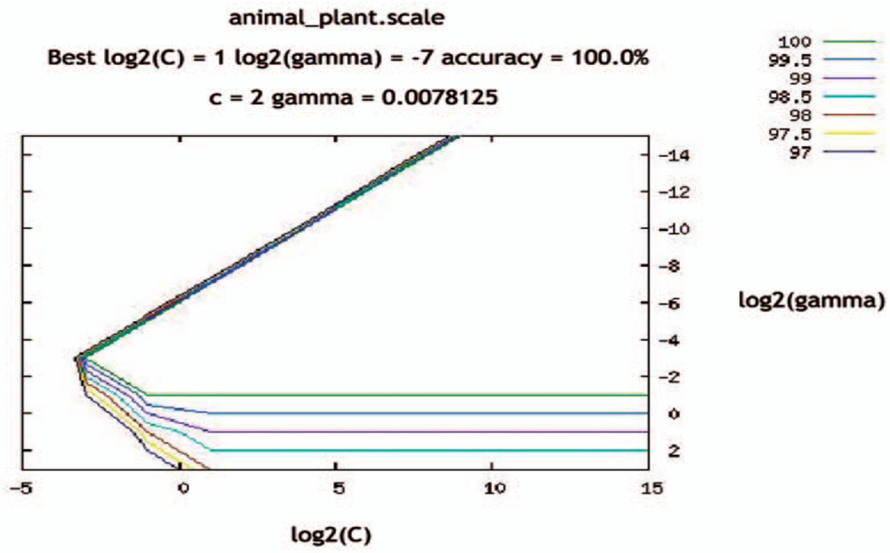
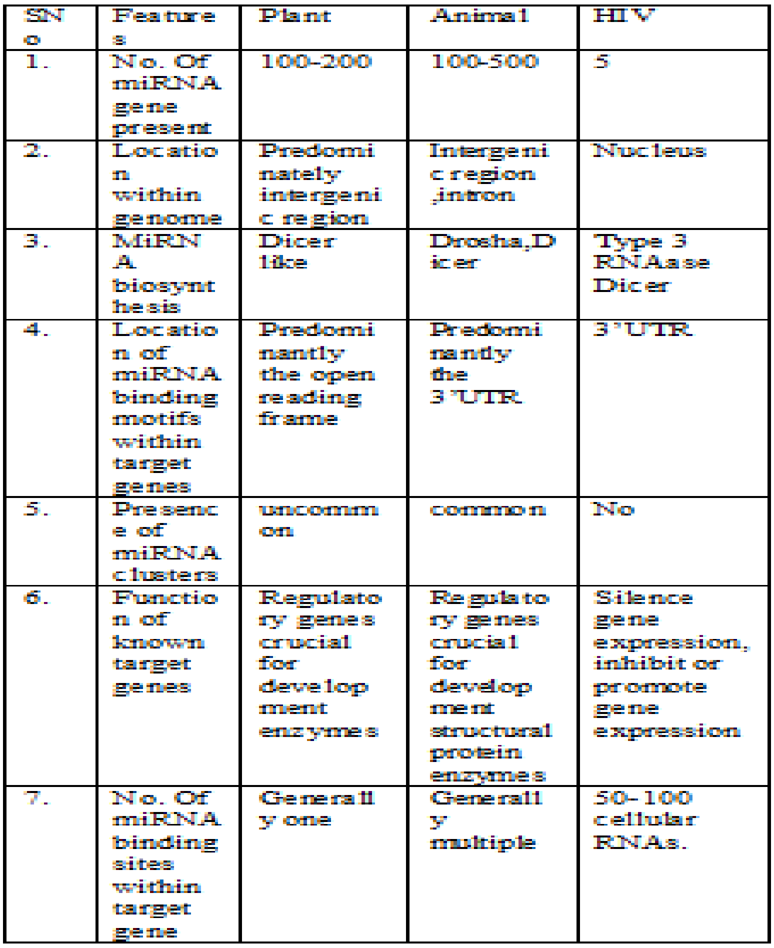


Fig. 2-Coarse grid search on C = 2-5, 2-4… 210 and gamma = 25,24

Table 1- shows dissimilarity between plant, animal and HIV as follows:



SNo	Features	Plant	Animal	HIV
1.	No. Of miRNA gene present	100-200	100-500	5
2.	Location within genome	Predominately intergenic region	Intergenic region ,intron	Nucleus
3.	MiRNA biosynthesis	Dicer like	Drosha,Dicer	Type 3 RNAase Dicer
4.	Location of miRNA binding motifs within target genes	Predominantly the open reading frame	Predominantly the 3'UTR	3'UTR
5.	Presence of miRNA clusters	uncommon	common	No
6.	Function of known target genes	Regulatory genes crucial for development enzymes	Regulatory genes crucial for development structural protein enzymes	Silence gene expression, inhibit or promote gene expression
7.	No. Of miRNA binding sites within target gene	Generally one	Generally multiple	50- 100 cellular RNAs.

Table 2- detailed accuracy by class

TP RATE	FP RATE	Precision	Recall	F-Measure	ROC Area	Class
0.988	0.222	0.953	0.988	0.97	0.883	Animal
0.778	0.012	0.933	0.778	0.848	0.883	Plant
0.95	0.184	0.949	0.95	0.948	0.883	Weighted average