Journal of Proteomics & Bioinformatics

Performance and Evaluation of MicroRNA Gene Identification Tools

Swati Sinha[1*], T.S. Vasulu[1], and Rajat K. De[2*]

[1]Biological Anthropology Unit, Indian Statistical Institute, Kolkata, India
[2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

*Corresponding authors: Swati Sinha, Biological Anthropology Unit, Indian Statistical Institute, Kolkata-108, India, Tel: 91-33-25753215; E-mail: swati.6783@gmail.com
Rajat K De, Machine Intelligence Unit, Indian Statistical Institute, Kolkata-108, India
Tel: 91-33-25753105, Fax (O): +91-33-25753026, E-mail: rajat@isical.ac.in

Received July 02, 2009; Accepted August 11, 2009; Published August 12, 2009

Citation: Sinha S, Vasulu TS, De RK (2009) Performance and Evaluation of MicroRNA Gene Identification Tools. J Proteomics Bioinform 2:336-343. doi:10.4172/jpb.1000093

Abstract

MicroRNAs are small single stranded RNA molecules of ~ 22 nt in length which play important role in post transcriptional gene regulation either by translational repression of mRNA or by their cleavage. Since their discovery, continuous efforts to identify the miRNA genes led to the discovery of several miRNAs in plants as well as animals. Owing to the limitations of the molecular genetic techniques of miRNA identification, computational approaches were introduced for better and affordable *in silico*-miRNA predictions. Here, we compared a few miRNA gene identification tools, such as 'MiPred','Triplet-SVM','BayesMiRNAfind','OneClassmiRNAfind'and 'BayesSVMmiRNAfind' to evaluate the performance of its predictability based on the real and pseudo precursor miRNA datasets. Of all the tools examined MiPred is more sensitive (96%) in identifying pseudo miRNAs than Triplet-SVM for real/pseudo miRNA classification, whereas for mature miRNA prediction 'one-class' SVM classifier shows best specificity (96%), while BayesSVMmiRNAfind shows least specificity (8%).

Keywords: MiPred; Triplet-SVM; BayesMiRNAfind; OneClassmiRNAfind; BayesSVMmiRNAfind; Sensitivity; Specificity; Accuracy; Mathew's correlation coefficient; Positive predictive value

Abbreviations: miRNA: MicroRNA; pre-miRNA: Precursor MicroRNA; HMM: Hidden Markov Model; SVM: Support Vector Machine; PCA: Principal Component Analysis; K-NN: K-Nearest Neighbor; MCC: Mathew's Correlation Coefficient; PPV: Positive Predictive Value

Introduction

Interest in miRNAs and their role as gene expression regulators has been growing immensely (Clop et al., 2006, Feng et al., 2009). The first effort that could identify such a small regulator, the lin-4 RNA in *C. elegans,* was done by Victor Ambros and colleagues, Rosalind Lee and Rhonda Feinbaum (Bartel, 2004). It was shown that the 21 nt lin-4 RNA, represses mRNA and controls part of the *C. elegans* larval development. The next small regulatory RNA to be discovered was the let-7, which controls another later developmental stage of *C. elegans* (Lee et al., 1993; Wightman et al., 1993). They were previously known as small temporal RNAs (stRNAs), but today recognized as the first of the large class of small regulatory non-coding RNA molecules, 'microRNAs'. Now it is believed that this class of molecules is not only limited to development but also plays a very important role in the regulation of a wide range of biological processes (Gard et al., 2006, Feng et al., 2009).

MicroRNAs are small non-coding RNAs of approximately 22nt (ranged 19-25nt) known to be involved in

posttranscriptional gene regulation either by cleavage of mRNAs or repressing the translation of mRNAs (Bartel DP, 2004). The microRNAs reported to be encoded within noncoding regions of genomes and within protein coding genes. The miRNA genes are transcribed by RNA polymerase II; in some cases RNA polymerase III may also be involved. Primary transcripts of the miRNA genes, 'pri-miRNAs', are processed in the nucleus to 'pre-miRNAs' by the RNAase III type endonuclease 'Drosha' and exported from the nucleus to the cytoplasm by nuclear export factor Exportin 5 and the Ran-GTP cofactor. The 60–90-nt miRNA precursors form the stem-loop structures, and the cytoplasmic ribonuclease class III enzyme 'Dicer' excises miRNAs from the pre-miRNA hairpin stem (Bartel, 2004). Dicer, either alone or with the help of Drosha, cleaves both strands of the precursor to form a double-stranded miRNA/miRNA* duplex (Dezulian et al., 2006). The mature miRNA strand of the duplex with the less stable 5' is then incorporated into RNA induced silencing complex (RISC) while the other strand is rapidly degraded.

Since the early studies elucidating the important role of miRNAs, there has been a continuous increase in the number of microRNAs along with simultaneous increase in the range of genomes encoding miRNAs, and it is worthy to mention that these small regulators play much more important role than previously thought (Gard et al., 2006). MicroRNAs and their associated proteins appear to be one of the more abundant ribo-nucleoprotein complexes in the cell. In general these have been detected by molecular genetic technique of cloning and Northern blotting but the detection of miRNAs whose expression is restricted to non-abundant cell types or specific environmental conditions could still be missed. Moreover these experimental methods are biased towards abundant microRNAs and are time-consuming. In these regard the computational approaches have been developed to complement experimental approaches to miRNA gene identification. It was found that the prediction of miRNA precursor candidates is fairly easy, whereas excluding false positives, as well as, exact prediction of the mature miRNA is a harder task. Thus, the computational detection methods must therefore be refined to serve as a faster, better, and more efficient method for miRNA detection.

In this regard, so far, several algorithms related to miRNA gene identification have been developed successfully. These approaches possess different roles like to predict the miRNA genes based on the evolutionary conservation of miRNA (miRScan, MirSeeker, MirFinder etc). Similarly various other bioinformatics' predictions combined with microarray analysis (PalGrade); HMM (ProMir) and structure-and-sequence analysis (miRAlign, MicroHarvester) helped to develop these different algorithms. Concept of machine learning approaches also was incorporated with the biology of miRNA to develop algorithms for the same purpose of miRNA gene identification, as for example: Diana-MicroH, mirCoS, miMatcher, Microprocessor-SVM, miRNA-SVM, RNAmicro, miREncoding etc. Apart from these algorithms, some other approaches do make predictions at different step of miRNA biogenesis e.g., some deal with the problem of classification, especially whether a given RNA sequence is miRNA precursor or not ('Triplet-SVM' and 'MiPred'); and some other approaches predict the location of mature miRNA in a given gene sequence ('BayesMiRNAfind', 'OneClassmiRNAfind', 'BayesSVMmiRNAfind' etc). Based on these two criteria, we have attempted to investigate the performance of the various tools of miRNA gene identification in this study. The comparative analysis can tell how well we can predict a given RNA sequence to be a precursor-miRNA and then how well we may know what is the location of mature miRNA in that particular precursor sequence.

Methodologies for Gene Identification

For the computational identification of miRNA gene various successful methods had been developed so far and it is a major thirst research area. This is because these tiny gene regulators have so many important functions such as their role in developmental timing, cell death, cell proliferation, haematopoesis and patterning of nervous system (Ambros, 2004). Lim et al., (2003) came up with an idea to identify miRNA genes that are conserved in more than one genome and developed miRScan for the same purpose. It was followed by another algorithm MirSeeker that analyzes the intronic and intergenic regions that are evolutionary conserved in *D. melanogaster* and *D. pseudoobscura* (Lai et al., 2003). The conservedness of these short sequences across species seems to be an important factor to develop an algorithm for miRNA gene prediction. Accordingly, Bonnet et al., (2004) developed MirFinder, a genome wide computational approach, to detect miRNA genes present in the *Arabidopsis* genome based on the conservation of sequences between *Arabidopsis* and *Oryza* genome. One more approach for the identification of miRNAs in *A. thaliana* was based on rigid complementarities that exist between plant miRNAs and their targets named as findMiRNA (Adai et al., 2005).

Apart from sequence conservation, other factors also helped to develop different algorithms for the same purpose. In 2005, Bentwitch et al. developed PalGrade, which

is an integrative approach that combines bioinformatics prediction with microarray analysis and sequence directed cloning. Information both at sequence and structural level helped to develop ProMir which is a probabilistic co-learning method based on paired HMM to identify close as well as distant homolog (Nam et al., 2005). Wang et al., (2005) were able to detect new miRNA based on structure and sequence alignment with a novel computational approach miRAlign. Similar sequence and structure information helped to develop MicroHarvester for the identification of candidate miRNA homolog in a set of sequences given a query miRNA (Dezulian et al., 2006). Another concept, that miRNAs are often found in clusters, was the basis of mirAbela (Sewer et al., 2005) that took into consideration only those genomic regions that are present around known miRNAs from mouse, human and rat.

Novel concepts kept on floating and led to novel discoveries, Berezikov et al., (2005) reported that phylogenetic shadowing of miRNAs in primate species revealed a characteristic conservation profile that can be used to detect the majority of known miRNAs efficiently and also predict an extensive novel set of miRNAs based on human-mouse-rat genome wide comparisons. On the other hand Lindow et al., (2005), developed 'microMatcher' for the identification of plant miRNAs that do not depend on phylogenetic conservation and identified 592 novel miRNAs which were not conserved in other plant genomes.

This is the present scenario where several prominent computerized miRNA detection approaches have been developed and utilized successfully. Most of these predictor algorithms depend on evolutionary conservation of miRNA sequences between different species. Such approaches allow filtering out many of the false-positive candidates, but they are obviously limited to detecting only the conserved microRNAs. Hence the concept of machine learning came into the picture when the need was to identify the non-conserved miRNAs. Such way of identification of novel microRNAs is a difficult pattern-recognition challenge. A single property is not sufficient for accurately detecting microRNAs, and in most cases rigid thresholds of the values for each of those properties is also not sufficiently sensitive. Rather, it is the combination of multiple properties, along with suitably different weighing of these different properties, that provides desirable accuracy. In this regard, many prediction algorithms based on machine learning have been developed. Xue et al., (2005), first developed an *ab initio* method called as Triplet-SVM, for distinguishing true pre-miRNAs from other pre-miRNA like hairpin structures taking into account a novel local contiguous structure sequence feature and used SVM with these features to classify real and pseudo pre-miRNAs. In order to improve the predictions by Triplet-SVM, another tool MiPred was developed which utilized the same feature (local contiguous structure sequence feature) in a hybrid way with another features including the MFE of the secondary structure, dinucleotide shuffling and P-Value of randomization test with a novel machine learning algorithm: Random Forest (Jiang et al., 2007). Several other algorithms made use of the SVM like Diana-MicroH (Szafranski et al., 2006) to predict miRNA hairpins using unique feature related to enzymatic cleavage with two additional features *viz.*, GC content and stem linearity; mirCoS (Sheng et al., 2007) that used three SVM models sequentially to discover novel miRNAs in mammalian genomes; Microprocessor-SVM that predicts processing sites for 50% of known human 5' miRNAs and miRNA-SVM that is trained on the output of the former one to identify non-conserved miRNAs (Helvik et al., 2007); RNAmicro, another SVM based approach that includes non stringent filter for consensus secondary structures and can easily identify pre-miRNAs in multiple sequence alignment (Hertel et al., 2006); miMatcher pipeline developed by Lindow et al.,(2007) performs intragenomic matching of potential miRNAs and their targets followed by classification of these miRNA candidates using SVM and miREncoding (Zheng et al., 2006) to encode the pre-miRNA sequences together with their secondary structures into the proposed 43 features using Weka software (Frank et al., 2004) in order to evaluate the performance of the selected classification algorithms along with the use of polynomial kernels for SVM algorithm. Apart from SVM Bayes classifier had also been used to develop new machine learning algorithm BayesMiRNAfind (Yousef et al., 2006), BayesSVMmiRNAfind (http://wotan.wistar.upenn.edu/BayesSVMmiRNAfind/) and OneClassMiRNAfind gene algorithm (Yousef et al., 2008). Brameier et al., (2007) used, for the first time, machine learning algorithm based on linear genetic programming and developed a unique *ab initio* method called as mirPRed for the prediction of novel mature miRNAs by genome scanning.

Materials and Methodology

Materials

The comparative study used five tools falling into two different categories. The first one includes 'MiPred' and 'Triplet-SVM' meant for the classification of real and pseudo pre-miRNAs and the second one includes

'BayesMiRNAfind', 'OneClassmiRNAfind',and 'BayesSVMmiRNAfind' for the prediction of mature miRNA. Out of the three tools in the latter category 'OneClassMiRNAfind' have options for five classifiers *viz.* SVM, Gaussian, Kmeans, PCA and K-NN and 'BayesSVMmiRNAfind' have options for two classifiers *viz.* naïve-Bayes and SVM. Positive as well as negative datasets were used for this comparative study. The positive dataset consists of 678 real miRNA precursor sequences of *Homo sapiens,* which were downloaded from the miRBase database release 11.0 (Griffiths-Jones et al., 2008, Griffiths-Jones et al., 2006; Griffiths-Jones, 2004). These sequences in the database have been either experimentally supported or obtained from literature mining and thus are the actual pre-miRNAs. The use of this positive dataset will therefore help to identify the number of true positives and false negatives that will define the sensitivity of a particular tool.

The negative dataset consists of 700 sequences of human pseudo pre-miRNAs. The source of this negative dataset is the coding dataset used by Xue et al., (2005). This 'coding dataset' consists of 8494 pre-miRNA like hairpins from which we collected the first 700 sequences (in order to make the number of positive and negative dataset sequences approximately equal) for this comparative analysis of various tools. These 700 sequences are the pre-miRNA like hairpins which are basically those sequence segments that have similar stem loop structure as actual miRNAs but still not been reported as pre-miRNAs (Xue et al., 2005). The use of this negative dataset will help to identify the number of false positives and true negatives that will define the specificity of a particular tool. Further details regarding the formation of the coding dataset can be obtained from Xue et al., (2005). The number of sequences given, as input, to all of these tools is same except for Triplet-SVM, where some sequences are not accepted due to the constraints in the algorithm.

Analysis

The positive and negative datasets are given as input to each of these tools and the output was analyzed to calculate the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

True positives is defined as the number of experimentally supported miRNA precursors that are predicted by a program and false negatives are those experimentally supported miRNA precursors that are not predicted by the program. Similar to the above case, false positives is defined as the number of all negatives that is predicted by a program and true negative on the other hand is all negatives that is not predicted by the program (Sethupathy et al., 2006).

Further, in order to evaluate the performance of these different predictive tools, we used the statistical parameters, *viz.,* Sensitivity (Se), Specificity (Sp), Accuracy (Acc),a summary statistic: Mathew correlation coefficient (MCC) and Positive predictive value (PPV). These parameters are based on TP, FN, TN and FP and are calculated using the following equations (Jiang et al., 2007):

Accuracy (Acc) = (TP+TN)/(TP+TN+FP+FN)*100
Specificity (Sp) = TN/(TN+FP)*100
Sensitivity (Se) = TP/(TP+FN)*100
MCC=((TP*TN)-(FP*FN))/
((TP+FP)*(TN+FN)*(TP+FN)*(TN+FP))^{½}

The Positive Predictive value (PPV) is calculated using the following formula (http://www.medicine.uiowa.edu/path_handbook/Appendix/Chem/ PRED_VALUE _THEORY.html).

Positive Predictive value (PPV) = TP/ (TP+FP) * 100.

a. Prediction of real/pseudo precursor-miRNA

Tools	Positive Data		Negative Data	
	TP	FN	FP	TN
MiPred	80.53	19.46	4.00	96.00
TripletSVM*	78.39	21.61	14.26	85.74

b. Prediction of mature miRNA

1. OneClassMiRNAfind				
SVM	71.24	28.76	5.57	94.43
Gaussian	69.62	30.38	13.43	86.57
K-means	78.91	21.09	13.43	86.57
PCA	73.16	26.84	13.43	86.57
KNN	80.09	19.91	13.43	86.57
2. BayesMiRNAfind	75.22	24.79	10.43	89.57
3. BayesSVMmiRNAfind				
Naïve-bayes	96.75	3.24	92.00	8.00
SVM	87.46	12.54	92.00	8.00

* The sample size: Positive (634), Negative (491)

Table 1: Percent count of TP, FN, FP and TN for two classes of predictions based on 678 real (Positive data set) and 700 pseudo (Negative data set) miRNA precursor sequences. a: Prediction of real/pseudo precursor miRNA, b: Prediction of mature miRNA.

Results

The results of the miRNA prediction tools are summarized in two ways; first in terms of percent count of TP, FN, FP and TN (Table 1) and second in terms of Sensitivity, Specificity, Accuracy, MCC and PPV of each tool (Table 2).

Table 1 shows the percent count of TP, FN, FP and TN obtained by the two types of predictive algorithms; a) Prediction of real/pseudo miRNA precursor and b) Prediction of mature miRNAs based on both positive and negative data sets.

a. Performance of real/pseudo precursor-miRNA prediction tools

Tools	Sp (%)	Se (%)	Acc (%)	MCC	PPV
MiPred	96.00	80.53	88.39	0.7761	95.12
TripletSVM	85.74	78.39	81.60	0.6362	87.65

b. Performance of mature miRNA prediction tools

1. OneClassMiRNAfind					
SVM	94.43	71.20	83.02	0.6768	92.53
PCA	86.57	73.15	79.97	0.6035	83.39
Kmeans	86.57	78.91	82.80	0.6572	85.06
Gaussian	86.57	69.61	78.23	0.571	84.07
K-NN	86.57	80.09	83.38	0.6684	85.24
2.BayesMiRNAfind	89.57	75.22	82.51	0.6556	87.47
3. BayesSVMmiRNAfind					
SVM	8.00	96.76	51.67	0.1029	50.46
Naïve Bayes	8.00	87.46	47.09	0.0853	47.94

Table 2: Percent values of the efficiency parameters: Specificity, Sensitivity, Accuracy. MCC and PPV to infer the performance of two predictive algorithms a: Performance of real/pseudo precursor-miRNA prediction tools, b: Performance of mature miRNA prediction tools.

Table 2 shows the performance evaluation indicators, especially Sensitivity, Specificity, Accuracy, MCC and PPV obtained by the same two types of predictive algorithms: a) Prediction of real/pseudo miRNA precursor and b) Prediction of mature miRNA.

Prediction of Real/pseudo miRNA Precursor

Of the two methods considered for classification of real and pseudo pre-miRNAs, MiPred is more sensitive in identifying the pseudo precursor miRNAs (96%), whereas TripletSVM is less accurate in identifying both types of miRNAs (Table 1a). On the other hand, the Specificity of MiPred is as high as 96% compared to the 85% specificity of Triplet-SVM. The results are in agreement to the total number of true negatives given by the respective tools. The Sensitivity of both these tools is almost same (78-80%) but because of the high specificity of MiPred the overall accuracy is more for this tool. Hence MiPred is a better tool when compared to Triplet-SVM which is also indicted by the 95.12% PPV for this tool. The MCC value for MiPred (0.7761) is high compared to the other tool TripletSVM (0.6362) showing the high efficiency of MiPred in classifying the real and pseudo pre-miRNA sequences (Table 2a).

Prediction of Mature miRNA

In case of predictive algorithms related to prediction of mature miRNAs, the two methods OneClassmiRNAfind and BayesMiRNAfind show similar results of correctly identifying the pseudo miRNA data (86-96%) than the real miRNA data (69-80%). Except the SVM classifier of OneClassmiRNAfind all other classifiers show relatively high percent of FP (13.43%) whereas the BayesMiRNAfind shows a lower FP value (10.43%), which is higher than that of one-class SVM classifier, but lower than all other one-class classifiers. Interestingly, the tool BayesSVMmiRNAfind was able to identify the real pre-miRNAs (87-96%) but is least efficient in identifying the pseudo pre-miRNA data with a TN value of 8 % only (Table 1b). On the other hand, the specificity of one class SVM classifier is best (96%) followed by BayesMiRNAfind (89.57%), whereas the specificities of all other one-class classifiers are in the same range (86.57%). Again one important observation is the very low specificity of BayesSVMmiRNAfind (8%) suggesting that this tool is not the preferred choice to correctly identify especially the pseudo pre-miRNAs. The sensitivities of the three tools along with their classifiers fall in the range of 71-87% except for BayesSVMmiRNAfind where it is 96%. This suggests that although BayesSVMmiRNA find is far from identifying the negative data correctly, but it can very well be used to predict the positive data. Further, the positive predictive value (PPV) of this tool is also low for both of its classifiers (for SVM it is 47.94% and for naïve Bayes it is 50.46%) when compared to other tools for the prediction of mature miRNA depicting the ineffectiveness of the predictions made by it. The overall accuracy of these tools is somewhat in the same range of 78-83% except for BayesSVMmiRNAfind, where there is a decrease in the range 47-51% because of its very low specificity. Further, the MCC value, which tells about the efficiency of the tool, falls in the same range (0.571-0.6768) but it is lowest in case of BayesSVMmiRNAfind for both of its

classifiers naïve-Bayes (0.0853) and SVM (0.1029) (Table 2b).

Discussion

The basic principle of many computational methods is to learn from known examples in order to find new ones and make better predictions. From a computational perspective, this corresponds to the problem of machine learning, an area of artificial intelligence used to develop techniques that allow computers to learn from examples. Since all the mechanisms behind miRNAs and their actions are not completely revealed, computational tasks associated with miRNA studies are often posed as a challenging machine learning problem with limited prior information (Yoon et al., 2006). In spite of such difficulties several algorithms based on the concept of machine learning have been developed but because the field of miRNA research is still in its blooming phase and the understanding at the molecular level is yet not very clear, the process of algorithms development for miRNA identification may not be completely exhaustive. It demands more understanding of the molecular aspect of miRNA biology and more clarity is needed to develop more accurate and efficient tools for the aforesaid purpose. Thus a comparative analysis of the performance of the various prediction tools available might be useful to carry out further research work in this area.

The present study takes into consideration the comparative analysis of the tools available for pre-miRNA classification and mature miRNA prediction in order to understand the limitations in these algorithms so that further efforts can be done for their improvement. Of all the tools studied, the earliest was TripletSVM, which included just the local contiguous sequence-structure feature. There is no inclusion of any thermodynamic related features hence there is a possibility of improvement in its overall performance. And it was achieved when MiPred was developed later on, which includes a hybrid feature incorporating local contiguous sequence structure feature along with the MFE and P value of randomization test. This improvement is visible in the result (Table 2) as the sensitivity as well as accuracy of MiPred is high compared to that of TripletSVM. Moreover, the specificity of MiPred is ~11% higher than TripletSVM; this is of much importance, as specificity is related to the false positive rate. More the specificity more will be the tendency of the program not to erroneously predict the negative data. The high performance of MiPred can also be contributed to the novel classifier algorithm Random Forest used in their program along with SVM.

Among the rest three tools, BayesMiRNAfind shows good specificity and accuracy indicating its better performance. The reason behind might be the inclusion of certain rules based on miRNA gene structure and sequence, thereby allowing prediction of non-conserved miRNAs. Moreover in order to reduce the false positive rate the tool is based on a comparative analysis over multiple species in an attempt to develop an algorithm that has higher specificity but almost similar sensitivity, which is also depicted by the results. OneClassMiRNAfind with SVM shows the best specificity among all classifiers which themselves have the same specificities. In case of sensitivities, OC-KNN and OC-K-means are superior to others as measured by their ability to capture only the known miRNAs. Two-class classifier approach BayesSVMmiRNAfind with SVM is showing the best sensitivity among all the tools compared but it has a very low specificity due to which the overall predictive criterion of accuracy is not good. Thus, BayesSVMmiRNAfind with SVM as a classifier has the highest sensitivity but its specificity is lowest (Figure 1).

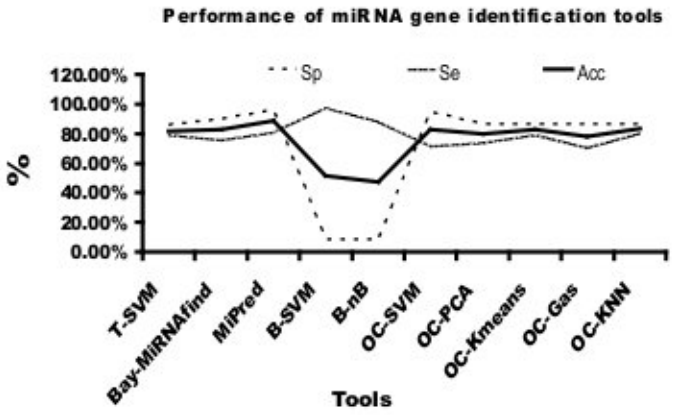

Figure 1: Comparison of miRNA gene identification tools as obtained from the evaluation measures of Specificity (Sp), Sensitivity (Se) and Accuracy (Acc).

Keeping in mind the various algorithms and methodologies developed so far one possible area of further research is to incorporate certain new features related to miRNA and to develop some new and more efficient algorithm for the same purpose. Furthermore, one possible limitation of the present study is that it is only based on comparing single programs and we have not considered the possibilities of combinations of several programs e.g., performance of various unions and intersections of individual programs, which might lead to a better comparative analysis.