Machine Learning methods for miRNA Gene prediction

Müşerref Duygu Saçar and Jens Allmer

Department of Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey

Key Words

Machine learning, miRNA gene prediction, miRNA gene detection, classification, test data, examples

Running Head

Machine Learning for miRNA prediction

Summary

MicroRNAs (miRNAs) are single-stranded, small, non-coding RNAs of about 22 nucleotides in length, which control gene expression at the posttranscriptional level through translational inhibition, degradation, adenylation, or via destabilization of their target mRNAs. Although hundreds of miRNAs have been identified in various species, many more may still remain unknown. Therefore, discovery of new miRNA genes is an important step for understanding miRNA mediated post transcriptional regulation mechanisms. It seems that biological approaches to identify miRNA genes might be limited in their ability to detect rare miRNAs and are further limited to the tissues examined and the developmental stage of the organism under examination. These limitations have led to the development of sophisticated computational approaches attempting to identify possible miRNAs *in silico*. In this chapter, we discuss computational problems in miRNA prediction studies and review some of the many machine learning methods that have been tried to address the issues.

1. Introduction

Current attempts to distinguish miRNA genes have led to the detection of thousands of miRNAs in various species, but many may remain undiscovered ***(1)***. These efforts, mainly based on experimental methods such as directional cloning of endogenous small RNAs, are time consuming, expensive, and work intensive ***(2)***. Inadequacy of experimental approaches can be showcased by the fact that miRNAs are expressed in specific cell types, at low levels or only in a specific condition which complicates their experimental detection. To overcome these problems several computational methods have been designed and applied to miRNA gene detection.

Numerous approaches for the *in silico* prediction of miRNAs have been created so far. These programs commonly regard the hairpin secondary structure of the miRNA precursor as the

most important characteristic of a miRNA gene ***(3, 4)***. RNA secondary structure prediction algorithms such as RNAfold ***(5)*** are used to predict the secondary structure and thermodynamic stability of RNA hairpin structures. Existing bioinformatics methods for the prediction of miRNA usually consist of: 1) genome-wide estimation of hairpin structures, 2) filtering or scoring of those hairpins based on their similarity in structure and sequence to known miRNA hairpins, and 3) experimental confirmation of putative candidates ***(3)***. In order to extract possible miRNAs from a genome, either homology modeling or *ab initio* methods are used.

2. Homology-based MicroRNA Gene Prediction

Homology-based miRNA gene mapping methods can build on available, experimentally validated, miRNAs and find similar structures and sequences in related species. The idea is that if a miRNA is identified in one genome then its homologs can be possibly found in other species ***(6)***. Since conservation indicates a function, it is assumed that conserved candidates are more likely to be miRNAs. Although it has been shown that for non-coding RNAs absence of conservation does not inevitably mean lack of function ***(7)***, searching for homologs especially in newly annotated genomes may be a beneficial approach. Software facilitating mapping of known miRNAs to homologous genomes take both sequence similarity and miRNA secondary structure information into account. The theory is based on derivation of mature miRNAs from hairpin structure formed by folding its pre-miRNA. The approach taken by one of the most recent developments, MapMi ***(8)***, first scans the miRNA sequences against the target genome and then creates potential pre-miRNAs from them. In the end, the results are scored, ranked and displayed. The scoring function considers both the quality of the sequence match (match, mismatch, perfect match) and the predicted structure of hairpins ***(8)***. The best candidate is chosen according to the calculated score and candidates are

further filtered based on a score-threshold which is either user-defined or selected from suggested thresholds ***(8)***. All candidates above threshold are displayed with their related scores and other relevant information. The web version of MapMi provides more detailed analysis including the generation and display of maximum likelihood phylogenetic trees, multiple sequence alignments and RNA Structural logos ***(8)***.

Although homology modeling can gather information from already successfully established miRNAs of a related organism's genome, it is also limited since completely novel miRNAs cannot be determined in this way. First attempts in this approach have mainly relied on identifying close homologs of published pre-miRs i.e. let-7 ***(9)***. This method might seem as straightforward as aligning sequences through NCBI BlastN ***(10)***, but it can only reproduce results and cannot find new miRNA genes. Since many miRNAs are species specific, they will always be missed by this method and therefore other strategies need to be used in tandem. Additionally, miRNA genes evolve very rapidly which further limits the applicability of homology-based methods ***(11)***. A powerful approach developed for genome-wide screening of phylogenetically well conserved pre-miRNAs between closely related species is cross-species sequence conservation based on computationally intensive multiple genome alignments. However, it also suffers lower sensitivity especially for more divergent evolutionary distances ***(12, 13)***. Moreover, identifying pre-miRNAs that differ significantly or undergo rapid evolution at the sequence level while keeping their characteristic evolutionary conserved hairpin structures, may also pose problems ***(2)***. Another important issue is that non-conserved pre-miRNAs with genus-specific patterns are likely to escape detection ***(2)***.

There are various homology-based miRNA gene prediction software such as MirScan ***(14)***, miROrtho ***(15)***, miRNAminer ***(16)***, and ProMiR II ***(17)***. Also, some of these software use machine learning approaches such as ProMir ***(18)***, which uses hidden Markov models, and MirFinder ***(19)***. MirFinder is designed for genome-wide, pair-wise sequences from two

chosen species and includes two key steps: 1) genome wide searching of hairpin candidates, 2) elimination of the non-robust structures based on 18 features analyzed by support vector machine (SVM) classification ***(19)***. The tool was tested on chicken/human, and *Drosophila melanogaster*/*D. pseudoobscura* pair-wise genome alignments. The results showed that the proposed method can be used for genome wide pre-miRNA predictions ***(19)***.

3. *Ab Initio*-based MicroRNA Gene Prediction

While homology-based methods mainly use comparative genomics, *ab initio* miRNA gene prediction needs no information other than the primary sequence for the prediction of miRNAs (albeit *ab inito* methods require negative and positive datasets; which is conceptually similar to homology-based approaches). Nonetheless, *ab initio* methods may enable the identification of new miRNAs which have no close homologs ***(20)***. The main difficulty of *ab initio* methods is choosing proper parameters that allow determining a given sequence to be a miRNA based on its properties. For instance, hairpin structure and minimum free energy are widely used features ***(21)*** of miRNAs in prediction tools such as miPred ***(2)***. If the chosen parameters do not provide good specificity, it would not be very informative and might increase the potential to produce false positive results. This would lead to a decrease in the accuracy of the miRNA prediction method and make validating the results of predictions in the wet-lab much more elaborate, time consuming and expensive. The main problem is that, although precursor miRNAs should possess an evolutionarily conserved hairpin structure which is critical for the early stages of the mature miRNA biogenesis, the hairpin shape is not unique to miRNAs and is found in many other non-coding RNAs ***(22)***. For instance, all translational RNAs contain multiple hairpins. It has been estimated that there are millions of hairpin like structures in the human genome and to differentiate the millions of hairpins from the few true miRNAs is the grant challenge ***(23)***.

There are many programs using *ab initio* methods with machine learning approaches including Triplet-SVM ***(24)***, MiRenSVM ***(22)***, miPred (SVM) ***(2)***, MiPred (Random Forest) ***(25)*** and MiRPara (SVM) ***(26)***.

4. Machine Learning and MicroRNA Gene Prediction

Next to defining proper features that allow differentiation between true and false miRNAs the selection of training data for machine learning algorithms is crucial for prediction success. Therefore, we will shortly comment on training and test data used in machine learning for miRNA prediction.

4.1. Learning and Test Data

4.1.1. Positive Data

Usually positive data for miRNA gene predictions are obtained from miRBase ***(27)***. However, there are some entries in miRBase which are suggested as miRNAs but are not fulfilling the necessary properties to be classified as miRNAs such as having more than one loop. It was shown that reference set of positive controls taken from miRBase, requires additional improvement to create a high-confidence set proper for use as positive controls ***(28)***. We recently elaborated on this and found that prediction accuracy can be improved upon filtering of unlikely miRNAs from miRBase ***(29)***. Except for these minor problems, in miRNA gene prediction studies, it is usually uncomplicated to select positive examples (e.g., using the known miRNAs), while it is challenging to create negative samples ***(6)***.

4.1.2. Negative Data

The collection of an appropriate negative dataset is vital for many machine learning algorithms to produce a well-trained classifier. If the sequences are too artificial, then there is a high probability that the machine learning method will not be trained adequately to

differentiate between true miRNAs and non-miRNA sequences ***(26)***. On the other hand, if the negative dataset is very similar to the positive dataset, the machine learning approach will be incapable of distinguishing between these two datasets ***(26)***.

One of the criteria for a small RNA sequence to be classified as a miRNA is that, it should be recognized and processed by the enzyme Dicer. While defining a negative dataset, this criterion should be used efficiently so that selected negative controls are not recognized by Dicer ***(28)***. The negative dataset sequences should be composed of transcripts that are expressed in the same cellular location as true miRNAs but are not recognized by Dicer. Since this is a very complicated way to generate negative samples, instead of this, in most of the algorithms random genomic sequences or exonic sequences are used ***(24, 28)***. These sequences are very weak negative controls because there is no confirmation that, these transcribed small RNAs would not be recognized by Dicer and other components of miRNA biogenesis pathway (i.e. Drosha, RISC) and processed into functional mature miRNAs ***(28)***. On the contrary, there is evidence that miRNAs can stem from any region of a genome (see other chapters in this volume) so that the assumption hairpins from exons are good negative data is quite dangerous.

A well-known negative dataset for miRNA gene prediction consists of 8494 pseudo hairpins from human RefSeq genes ***(30)*** which have been selected such that they do not undergo any alternative splicing events ***(2)***.

4.2. Algorithms for Machine Learning

Machine learning is used in many bioinformatics applications and studies (Figure 1). The quickly increasing amount of data, created by modern molecular biology techniques, has caused the need for accurate classification and prediction algorithms since handling it with traditional methods is not feasible anymore ***(31)***. There are numerous biological fields where

machine learning methods are applied for knowledge extraction from data such as genomics, system biology, evolution, microarray, proteomics ***(32)***.

Machine learning algorithms are different from the rule-based miRNA prediction algorithms since the rules to decide whether a given sequence is a miRNA, are not manually created, instead these rules are fit, trained, or learned from examples ***(32)***. Usually machine learning-based methods start with the learning process of sequence, structure or thermodynamic characteristics of miRNAs. Next, a classifier is formed to decide whether unknown sequences are true miRNAs based on the information gained through positive and negative data sets. Normally, the parameters are a set of numerical features defining a candidate miRNAs such as minimum free energy of folding and the results would be true or false indicating whether the candidate is a miRNA or not.

However, there are two main weaknesses with the existing machine learning based miRNAs gene identification methods. The first one is the imbalance between positive and negative examples. Since the exact number of real miRNAs in any genome is unknown, it is supposed that there are few miRNA precursors in a genome ***(22)*** so that any arbitrarily selected hairpin extracted from the genome is unlikely to be a pre-miRNA. Also, the number of positive examples is significantly smaller than that of generated negative examples (note caution in 4.1.2). For instance, one of the commonly used negative dataset for miRNA prediction algorithms consists of approximately 9000 pseudo hairpins while the number of human miRNAs that can be obtained from miRBase is less than 1500 ***(2)***. The imbalance problem between positive and negative datasets can significantly reduce the performance of current machine learning approaches ***(22)***. The other problem is that most of the current machine learning based algorithms make assumptions such as length of the stem, loop size and minimum free energy (MFE) of the data. Thus, sequences outside of these predetermined

borders are not considered as a true miRNA and cannot be predicted by those methods, which may reduce the prediction performance and accuracy ***(22)***.

To our knowledge, there is no published study that uses unsupervised machine learning approaches for miRNA gene prediction. On the other hand, there are many studies using supervised machine learning algorithms such as support vector machine (SVM), neural networks (NN), hidden Markov models (HMM), and Naive Bayes (NB) (for more details on these algorithms see **Chapter 7** in this volume).

4.2.1. Supervised MicroRNA Gene Prediction Approaches (Classification)

Machine learning for miRNA gene prediction is almost exclusively based on supervised learning in which an algorithm is trained to learn; approximating a function that maps input data to required outputs ***(33)***. Usually, the inputs are a set of parameters designating a candidate (e.g., mfe, number of dinucleotides, length of stem, etc.) and the output would be miRNA or non-miRNA. While the anticipated output is unknown, the machine is trained by input. The main idea of the process is that the machine learner should be capable of simplifying from these examples (input data; positive and negative examples) and properly classify candidates ***(6)***. The most important factor influencing the accuracy of the results is the choice of features since parameterization of the examples into features is not performed automatically ***(6, 22)***. To test the accuracy and precision of the machine learning process, a system called cross-validation is used. Cross-validation is important to prevent Type III errors (as put by Mosteller: "*correctly rejecting the null hypothesis for the wrong reason*" ***(34)***), particularly in situations where further samples are dangerous or expensive to obtain. One round of cross-validation includes dividing a sample of data into corresponding subsets, performing the analysis on one subset (the training or learning set), and validating the analysis on the test set (Figure 2). The example sets can be divided in defined percentages (e.g. 70% of samples included in learning set, remaining 30% included in testing set. See Figure 2) but the

essential point is that these datasets must not have shared examples. After cross-validation the best model is selected and applied to perform predictions.

One of the initial works in the field by Sewer et al. (2005) assembled 40 different sequence and structural parameters to label a candidate as pre-miRNA. The SVM classifier model was trained using 178 known human pre-miRNAs as positive examples and 5395 random sequences obtained from tRNA, rRNA, and mRNA genes as negative examples (in reality, there is no guarantee that these RNAs would not contain any functional miRNAs, see Section **4.1.2**). As a result of huge difference between the number of positive and negative samples, their results have high specificity (91%) and low sensitivity (71%) for their dataset.

ProMiR was introduced in 2005 as an algorithm that uses a Hidden Markov Model and simultaneously takes into account structure and sequences of pre-miRNAs (Nam et. al. 2005). A machine learning approach was used with positive examples from known human miRNAs and negative examples obtained arbitrarily from the human genome. The predicted pre-miRNAs are further assessed according to their minimum free energy and searched to find out whether they are conserved among vertebrates. ProMiR II includes additional features than ProMiR such as addition of knowledge about miRNA gene clustering, G/C ratio conservation, and entropy of candidate sequences (Nam et. at. 2006).

MatureBayes is a probabilistic algorithm developed by Gkirtzou et. al., which uses a Naive Bayes classifier to characterize potential mature miRNAs ***(35)***. Similar to previous approaches, it also performs classification based on sequence and secondary structure information of miRNA precursors.

4.2.2. One-class Classification

The major challenge of classification is appointing a new object to one of a set of classes which are defined in advance. This classification process is performed by using the learned

rules based on a number of examples. Differing from other classification approaches, in one-class classification it is supposed that only information of one of the classes, also known as the target class, is accessible. Hence, since there is no information apart from the examples of the target class, the distinction between the two classes has to be assessed from data of only the real class ***(36)***.

Defining the negative class is the most difficult challenge to overcome in developing machine-learning algorithms for miRNA identification. Therefore, machine-learning approaches have been proposed for identifying miRNAs without the requirement of a negative class. Yousef and colleagues performed a study using one-class machine learning approach for miRNA gene prediction by using only positive data to construct the classifier ***(37)***. Although one-class method is less complex to implement which makes it easier to handle, the two-class procedures generally seem to be superior. Moreover, there are additional problems due to some characteristic properties of miRNAs; e.g. pre-miRNAs must fold in a hairpin structure but not all the hairpins in the genome are miRNA sequences ***(38)***.

5. Conclusion

The biggest challenge for miRNA gene prediction is that most eukaryotic genomes include vast numbers of inverted repeats (IR) so the transcripts of these IRs can form strong hairpins ***(6)***. Without considering phylogenetic conservation it has been shown that about ≈11 million hairpins can be found in the human genome ***(1)***. These hairpins can have various origins and take part in numerous processes one of which might be miRNA mediated posttranscriptional regulation ***(6)***. Since not all hairpins are miRNAs, identifying the hairpins which would become functional miRNAs is a very difficult task. Moreover, the big number of possible hairpins makes reducing the false positive rate and increasing the accuracy of the prediction a difficult task.

Machine learning approaches have become popular for miRNA gene prediction studies. Since there are known miRNAs either experimentally validated or discovered through bioinformatics tools, positive datasets which is a necessity for machine learning methods, are available for miRNA precursors. Moreover, there are also some rules defining a sequence as a miRNA (e.g. recognition and being processed by miRNA biogenesis pathway enzymes such as Dicer and Drosha) so the sequences that do not pass this criteria can be used as negative datasets. However, it is important to keep in mind that the quality of these datasets will affect the sensitivity and specificity of the designed programs (see Section **4.1**). Still, in order to overcome the difficult issue of creating appropriate negative datasets one-class classification method can be applied to the miRNA gene prediction problem. The abundance of machine learning methods employed for miRNA gene prediction shows that these approaches are deemed to be suitable to deal with this problem.

6. Acknowledgements

This study was in part supported by an award received from the Turkish Academy of Sciences for outstanding young scientists (TUBA GEBIP, http://www.tuba.gov.tr).

7. References

1. Bentwich, I., Avniel, A., Karov, Y., et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics* **37**, 766–70.

2. Ng, K.L.S. and Mishra, S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics* **23**, 1321–30.

3. Burgt, A. van der, Fiers, M.W.J.E., Nap, J.-P., et al. (2009) In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC genomics* **10**, 204.

4. Janssen, S., Schudoma, C., Steger, G., et al. (2011) Lost in folding space? Comparing four variants of the thermodynamic model for RNA secondary structure prediction. *BMC Bioinformatics* **12**, 429.

5. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431.

6. Lindow, M. and Gorodkin, J. (2007) Principles and limitations of computational microRNA gene and target finding. *DNA and cell biology* **26**, 339–51.

7. Pang, K.C., Frith, M.C., and Mattick, J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in genetics : TIG* **22**, 1–5.

8. Guerra-Assunção, J.A. and Enright, A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**, 133.

9. Pasquinelli, A.E., Reinhart, B.J., Slack, F., et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–9.

10. McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research* **32**, W20–5.

11. Liang, H. and Li, W.-H. (2009) Lowly expressed human microRNA genes evolve rapidly. *Molecular biology and evolution* **26**, 1195–8.

12. Berezikov, E., Guryev, V., Belt, J. van de, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–4.

13. Boffelli, D., McAuliffe, J., Ovcharenko, D., et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science (New York, N.Y.)* **299**, 1391–4.

14. Lim, L.P., Lau, N.C., Weinstein, E.G., et al. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev* **17**, 991–1008.

15. Gerlach, D., Kriventseva, E. V, Rahman, N., et al. (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.* **37**, D111–117.

16. Artzi, S., Kiezun, A., and Shomron, N. (2008) miRNAminer: A tool for homologous microRNA gene search. *BMC Bioinformatics* **9**, 39.

17. Nam, J.-W., Kim, J., Kim, S.-K., et al. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.* **34**, W455–458.

18. Nam, J.-W., Shin, K.-R., Han, J., et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* **33**, 3570–3581.

19. Huang, T.-H., Fan, B., Rothschild, M.F., et al. (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics* **8**, 341.

20. Brameier, M. and Wiuf, C. (2007) Ab initio identification of human microRNAs based on structure motifs. *BMC Bioinformatics* **8**, 478.

21. Allmer, J. and Yousef, M. (2012) Computational methods for ab initio detection of microRNAs. *Frontiers in genetics* **3**, 209.

22. Ding, J., Zhou, S., and Guan, J. (2010) MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC Bioinformatics* **11 Suppl 1**, S11.

23. Bentwich, I. (2008) Identifying human microRNAs. *Current Topics In Microbiology And Immunology* **320**, 257–69.

24. Xue, C., Li, F., He, T., et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6**, 310.

25. Jiang, P., Wu, H., Wang, W., et al. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35**, W339–344.

26. Wu, Y., Wei, B., Liu, H., et al. (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics* **12**, 107.

27. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* **39**, D152–7.

28. Ritchie, W., Gao, D., and Rasko, J.E.J. (2012) Defining and providing robust controls for microRNA prediction. *Bioinformatics (Oxford, England)* **28**, 1058–61.

29. Saçar, M.D., Hamzeiy, H., and Allmer, J. (2013) Can MiRBase Provide Positive Data for Machine Learning for the Detection of MiRNA Hairpins? *Integrative Bioinformatics* **(accepted)**,.

30. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–4.

31. Bhaskar, H., Hoyle, D.C., and Singh, S. (2006) Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Computers in biology and medicine* **36**, 1104–25.

32. Larrañaga, P., Calvo, B., Santana, R., et al. (2006) Machine learning in bioinformatics. *BRIEF BIOINFORM* **7**, 86–112.

33. Zhang, Y.-Q., Rajapakse, J.C., Zhang, B.-T., et al. (2008) Supervised Learning Methods for MicroRNA Studies., *Machine Learning in Bioinformatics*, p. 339 John Wiley & Sons, Inc.

34. Mosteller, F. (1948) A k-Sample Slippage Test for an Extreme Population. *The Annals of Mathematical Statistics* **19**, 58–65.

35. Gkirtzou, K., Tsamardinos, I., Tsakalides, P., et al. (2010) MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. *PloS one* **5**, e11843.

36. Tax, D.M.J. (2001), One-class classification., ISBN: 90-75691-05-x.

37. Yousef, M., Jung, S., Showe, L.C., et al. (2008) Learning from positive examples when the negative class is undetermined--microRNA gene identification. *Algorithms for molecular biology* **3**, 2.

38. Bentwich, I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett.* **579**, 5904–5910.

Figure Captions

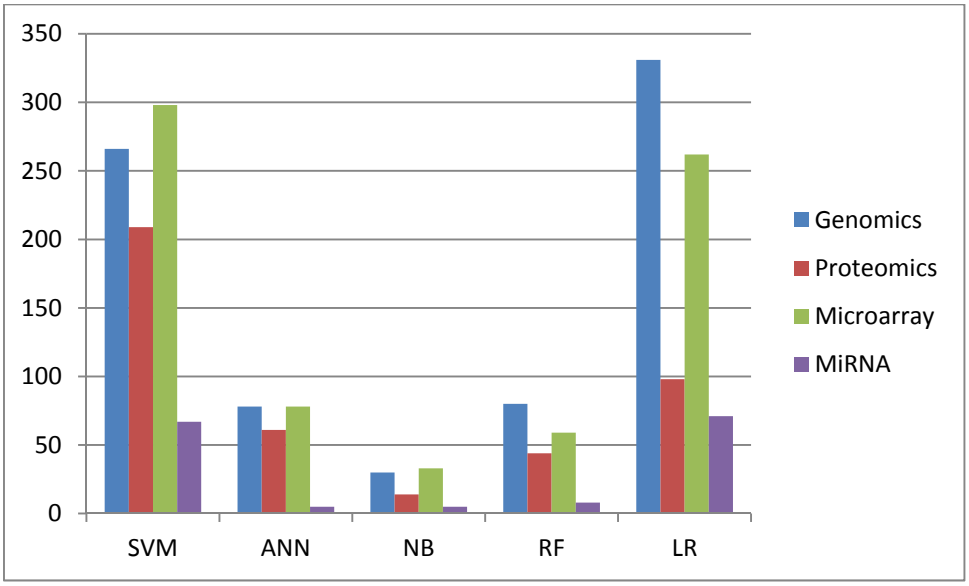

Figure 1. Fields in biology where machine learning methods are applied (The number of publications (y-axis) is calculated by searching PubMed with machine learning approach and the field name as key words).

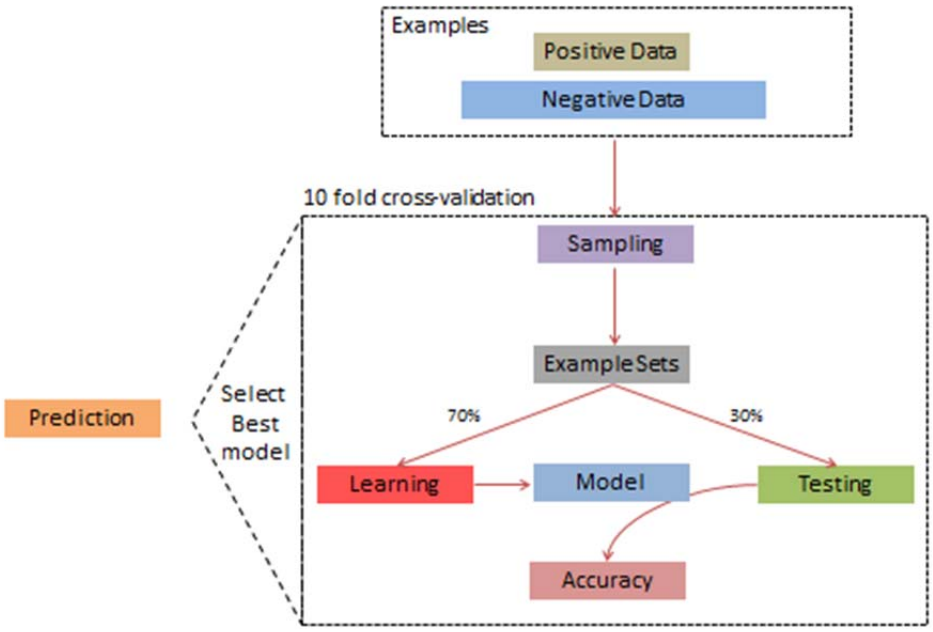

Figure 2. General work-flow of machine learning algorithms for miRNA gene prediction.