Journal of Biomedical Informatics 46 (2013) 563–573

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role

Dimitrios Kleftogiannis[a], Aigli Korfiati[b], Konstantinos Theofilatos[b,*], Spiros Likothanassis[b], Athanasios Tsakalidis[b], Seferina Mavroudi[b,c]

[a] King Abdullah University of Science and Technology (KAUST), Computer Science and Mathematical Sciences and Engineering Division, Thuwal, Saudi Arabia
[b] Department of Computer Engineering and Informatics, School of Engineering, University of Patras, Greece
[c] Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Patras, Greece

ARTICLE INFO

Article history:
Received 13 October 2012
Accepted 12 February 2013
Available online 8 March 2013

Keywords:
microRNA genes
miRNA identification
miRNA gene prediction
Gene regulatory networks
miRNA transcription mechanism
Machine learning

ABSTRACT

Traditional biology was forced to restate some of its principles when the microRNA (miRNA) genes and their regulatory role were firstly discovered. Typically, miRNAs are small non-coding RNA molecules which have the ability to bind to the 3′untraslated region (UTR) of their mRNA target genes for cleavage or translational repression. Existing experimental techniques for their identification and the prediction of the target genes share some important limitations such as low coverage, time consuming experiments and high cost reagents. Hence, many computational methods have been proposed for these tasks to overcome these limitations. Recently, many researchers emphasized on the development of computational approaches to predict the participation of miRNA genes in regulatory networks and to analyze their transcription mechanisms. All these approaches have certain advantages and disadvantages which are going to be described in the present survey. Our work is differentiated from existing review papers by updating the methodologies list and emphasizing on the computational issues that arise from the miRNA data analysis. Furthermore, in the present survey, the various miRNA data analysis steps are treated as an integrated procedure whose aims and scope is to uncover the regulatory role and mechanisms of the miRNA genes. This integrated view of the miRNA data analysis steps may be extremely useful for all researchers even if they work on just a single step.

1. Introduction

Scientific findings in the previous decade indicated that cells contain a variety of non-coding RNAs including components of the regulatory mechanisms that govern most of the cellular procedures [1]. The development of high throughput techniques accelerated the discovery of such small non-protein-coding RNAs. MicroRNAs (miRNAs) are non-coding regulatory molecules which scientific evidence has proved to be small (approximately 22 nt) in length, stable and to regulate the functions of many other target genes. Typically, miRNAs have the potential to bind to the 3′untraslated region (UTR) of their mRNA target genes for cleavage or translational repression [2]. The discovery of the miRNA genes and their targets made a breakthrough in the molecular diagnostics and consequently the better understanding of the underlying mechanisms may lead to more sophisticated therapeutic strategies.

The very first miRNAs and their targets were discovered experimentally through classical genetic techniques. A description of the experimental techniques and the detailed history of the miRNA genes discovery can be found in [2]. However, the experimental identification of miRNA genes and their targets has many drawbacks such as expensive experimental protocols (cost), time consuming experiments and low specificity. This is the reason why computational techniques have been proposed to overcome these drawbacks.

Computational approaches for the identification of miRNA genes are classified in comparative and non-comparative ones [3,4]. Comparative methods, which were the first computational approaches for the identification of miRNAs, are based on the conservation of the genomes in close species. Unfortunately, these methods are not able to predict non-conserved miRNAs. In opposite, non-comparative techniques are based on conservative features, and deploy machine learning (ML) techniques to predict miRNAs using a variety of sequential, thermodynamical and structural features.

One other important problem, relevant to the analysis of miRNAs, for which computational approaches are required, is the problem of predicting miRNA target genes [3,4]. From a computer science perspective many different approaches have been developed and they are divided in filtering approaches and ML

[*] Corresponding author. Address: Building B, University Campus Rio, Patras, Greece. Fax: +30 2610338798.
E-mail address: theofilk@ceid.upatras.gr (K. Theofilatos).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.02.002

approaches. Filtering approaches filter an initial large set of candidate miRNA target genes using statistical measures and various conservation, sequence matching, structural and thermodynamical criteria in order to end up with a small set of target candidates. ML approaches, which are believed to overwhelm the filtering ones in terms of prediction performance, deploy techniques such as Naïve Bayesian Classifiers, Artificial Neural Networks and Support Vector Machines (SVMs), using as inputs various features from different categories which are calculated for each miRNA and candidate target gene pair.

Lately, the adequate performance that has been achieved for the aforementioned miRNA analysis problems has enabled researchers to study the transcription mechanisms of miRNA genes [5] and their participation in gene regulatory networks [6]. Some computational techniques have been developed for these tasks and their first results are extremely promising. The ultimate goal is to uncover the regulatory role of every miRNA and to distribute this information publicly.

All the computational intelligence techniques for the analysis of miRNA data, present certain advantages and disadvantages. When predictors–classifiers need to be trained, many problems arise concerning the trade-off between sensitivity and specificity, the creation of the negative training set, overfitting issues, the selection of the appropriate inputs to be used and class imbalance problems. All these issues and some additionally ones concerning the accessibility of the existing tools and the efficient and fair comparison of the computational techniques are discussed in the present survey which aims to provide useful guidelines for researchers which are going to work on any of the examined miRNA data analysis tasks. Moreover, some interesting future directions are proposed.

The rest of the present paper is organized as follows: Section 2 reviews the existing methodologies for the identification of miRNA genes, while Section 3 presents and compares existing techniques for the prediction of miRNA target genes. Section 4 emphasizes on computational techniques for the analysis of the miRNA transcriptional mechanism and Section 5 examines methodologies for the construction of miRNA enriched gene regulatory networks. Finally, Section 6 discusses the most important issues for the computational analysis of miRNA data and proposes some interesting future directions.

2. microRNA genes prediction

Three major in vivo procedures have been applied for the experimental verification of microRNA candidates: Northern blot, in situ hybridization and Polymerase Chain Reaction (PCR) assays [7]. The short length of microRNA genes, their ability to act redundantly and the low expression profiles are the major limitations of the experimental techniques. Consequently all these procedures are not suitable for the prediction of novel miRNAs and they can be used for the experimental verification of miRNA candidates [8].

In order to overcome the technical hurdles and the limitations of the experimental techniques many computational approaches have been developed. Previous review papers have already tried to report the advantages and disadvantages of the existing methodologies. Zhang et al. [3] described the major characteristics of miRNAs and reported the available methods for gene and target prediction. Also they reported tools for miRNA secondary structure analysis and miRNA databases such as the miRBase and miRNAMap. The differences in identification between plants and animals are described in [4]. The state of the art methodologies and their limitations were addressed by Lindow and Gorodkin in [9]. Specifically, they emphasized on the information which can be exploited by the data and they addressed the importance of developing more general frameworks for the prediction of both genes and targets. According to the applied training data sets Mendes et al. in [10] separated the available methods in animal based and plant based. Then, they sub-divided them in different categories such as filter based, machine learning, homology search, and target-centered. They concluded with the central problems of this research area and they reported the needs for the selection of more discriminative features, the development of better models for the miRNA-like pseudo hairpins and the better understanding of the miRNA biogenesis mechanism. Pseudo hairpins are hairpin sequences with similar to miRNAs stem-loop features. Similar to the previous works, Li et al. [11], in contrast to Mendes et al., separated the available methods in three categories according to the conservation criteria, the ML algorithms and the experimental data. Concerning the future work, they suggested the integration of miRNA regulatory information and they characterized as an emerging issue the discovery of the miRNA functional binding sites.

In this review paper in order to give a simplified overview of the existing methodologies and to classify correctly the large number of mixed approaches we categorized the available methods into two basic categories: the comparative or conservation-based approaches and the non-comparative methods or ML approaches. In addition, the present advances in high throughout techniques boosted the development of methods which integrate different types of data and perform high level analysis.

The earliest approaches for discovering pre-miRNAs are based on comparative techniques and they can identify miRNAs with close homologs among species [12]. They rely on strong conservation criteria within and between species. The basic idea is to use comparative genomics to filter out hairpins that are not evolutionarily conserved in related species. Advanced comparative techniques rely on sequence similarity and the most straightforward method is to align unknown RNA sequences to known pre-miRNAs with a BLAST-like algorithm. On the other hand the non-comparative methods do not rely on the heavily phylogenetic conservation. They emphasize on ML algorithms to scan for miRNA candidates. In order to distinguish real miRNAs from other secondary structures that fold into a similar hairpin structure (pseudo hairpins), various classification systems have been developed. In every ML method the first step consists of the computation of topological, sequential, thermodynamical and other characteristics of the miRNAs. Then a classifier is trained based on these features and a model that is capable of predicting candidate miRNAs is constructed. Various ML algorithms have been proposed and the most important of them are going to be described next. In Table 1 we present the most important computational methods for the identification of miRNAs which may be accessed through a web interface.

2.1. Conservation-based approaches

One of the first conservation-based approaches was **MiRScan** [13] which relied on the observation that known miRNAs are derived from phylogenetically conserved stem loop precursor RNAs. The program was applied to known conserved patterns of *Caenorhabditis elegans* and successfully predicted close homologs of *Caenorhabditis briggsae*. Likewise the **MiRseeker** [14] method was applied to the euchromatic sequences of *Drosophila melanogaster* and *Drosophila pseudoobscura*. Conserved sequences that adopt an extended stem-loop structure were analyzed. The method correctly identified most of the previously identified Drosophila miRNAs, but it presented low sensitivity.

The **findMiRNA** method [15] was developed in order to detect miRNAs of the *Arabidopsis thaliana* genome. This algorithm relied on the more rigid complementarity between existing plant miRNAs and their targets to identify initial miRNA candidates. The main contribution of the technique was the usage of known targets to filter out and predict new miRNAgenes. The **microHARVESTER** methodology [16] by taking advantage of the state of the art RNA analysis techniques tried to detect miRNA genes based on conservation patterns and sequence similarity criteria. The **MirAlign** program [17] focused on the genome wide detection of animal miRNAs. The algorithm was based on structure and sequence alignment and the approach achieved higher performance than the other reported comparative methods.

The **MirCheck** method [18] focused on the identification of both miRNAs and targets of *A. thaliana* and *Oryza sativa* genomes. The contribution of the method was of great importance and established new significant biological knowledge regarding the miRNA class and their regulatory functions.

All the reported comparative methods failed to predict miRNAs that do not have clear homologs among species. On the other hand they established new knowledge for the microRNA class and increased the number of identified miRNA genes by experimental verification techniques. The main drawback is the low sensitivity in divergent evolutionary distance [19]. Moreover non-conserved miRNAs with genus-specific patterns are likely to evade detection. Recent reports showed that the number of non-conserved miRNAs missed by the comparative techniques was still large. The viral miRNAs are striking examples of that category [20]. Also, the unreliability of alignment algorithms creates limitations and further reduces the capability of comparative methods in identifying novel miRNAs.

2.2. ML approaches

In order to overcome the limitations of the conservation based techniques and increase the prediction performance, many research groups applied more sophisticated algorithms to classify miRNA candidates. Among the different methodologies the SVM

Table 1
Computational tools for the identification of miRNAs.

Method	Category	Advantages	Disadvantages	Website
MiRScan	Conservation based	High predictive performance for miRNAs similar to existing homologs	Low prediction performance in no-close homologs	http://genes.mit.edu/mirscan/
MirCheck	Conservation analysis	Prediction and experimental verification of miRNA genes that had not previously been identified Significant biological knowledge	No thermodynamic information Limited predicting performance	http://web.wi.mit.edu/bartel/pub/software.html
MirAlign	Comparative analysis	Higher performance than the other homologous searching approaches	Low sensitivity	http://bioinfo.au.tsinghua.edu.cn/miralign
BayesMirFind	Machine learning	Large and representative feature set Try to reduce false positive rate	Very simple classification method No feature selection method	https://bioinfo.wistar.upenn.edu/miRNA/
RNAmicro	Machine learning	Promising efficiency The small feature set has discriminative power	The performance depends on the sensitivity and specificity of the initial screen for structured RNA candidates	http://www.tbi.univie.ac.at/~jana/software/RNAmicro.html
miPred	Machine learning	Applied in human, no human, other non-coding and viral datasets Representative feature set containing various types of information Studies the contribution of individual features	Lack of a functional web server The tuning of the parameters needs improvements	http://web.bii.a-star.edu.sg/archive/stanley/Publications/
MiRFinder	Machine learning	Alternative approach to the standard comparative techniques which combines genome wide scan with ML	Low performance	http://www.bioinformatics.org/mirfinder/
Theofilatos et al.	Machine learning	Systematic approach for simultaneous feature selection and SVM parameters optimization Interpretable Results High classification metrics	The method did not apply on various species Did not include new features High computational complexity	http://150.140.142.24:82/Default.aspx
miRD	Machine learning	Tested on NGS datasets Functional web server Representative datasets	Some variations of single stem and multi stem strategies achieve low performance The construction of training and test sets is not described adequately	http://mcg.ustc.edu.cn/rpg/mird/upload.php
Xiao et al.	Machine learning	Studies the contribution of individual parameters (deleting one by one) Relatively big training set	Comparable to other methods the predicting performance is limited	http://cic.scu.edu.cn/bioinformatics/Pre-miRNA_code.zip
miRDeep	ML and NGS data	Very promising novel approach Flexible tool to analyze deep sequencing data	The miRDeep scoring algorithm models the miRNAbiogenis, an open-research and partially unknown procedure Parameter estimation	http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html
miRanalyzer	ML and NGS data	Web server for NGS analysis that integrates miRNA detection and prediction	No thermodynamic characteristics and the very simple classification algorithm implementation	http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php
MiREna	Comparative/computational	Performance similar to those of ML approaches It can be applied on different types of data	No representative criteria The performance is comparable to other ML classifiers	http://www.ihes.fr/%CB%9Ccarbone/data8/

classifier was used more frequently. Also Hidden Markov Models (HMMs), Naïve Bayes classifiers and lately Random Forest models were proposed. The rest of this section will present the most important ML gene prediction techniques grouped by the type of their classifier.

2.2.1. Support Vector Machine classifiers

In 2005, the **TripletSVM** method [21] introduced a set of local contiguous structure-sequence features and developed a SVM classifier to distinguish real miRNAs from pseudo hairpins. This method achieved 90% accuracy in the human dataset and lower performance in plants and viruses. The **mirAbela** [22] method focused on genomic regions of known human, mouse and rat miRNAs. The method was capable of discovering clustered and probably co-transcribed, miRNAs. However, it achieved low specificity and sensitivity. In order to overcome the drawbacks of the previous methods, **Diana-MicroH** [23] was introduced as a tool for predicting miRNAs with high predicting performance. The authors implemented a SVM classifier trained on structural and evolutionary features of miRNA hairpins. The method also incorporated information on how the enzymatic cleavage occurs and studied the discriminative power of several attributes.

The **RNAmicro** [24] introduced an SVM approach which was capable of recognizing microRNA precursors using 12 features based on structure, sequence composition, sequence conservation, thermodynamic stability and structure conservation. The usage of that small feature set showed promising efficiency, but the method's performance was correlated to the structure of the initial miRNA candidates. Among the SVM approaches the contribution of the **miPred** method [25] was of great importance. The authors applied a SVM classifier and they used representative global and intrinsic folding attributes without relying on phylogenetic conservation. They collected a feature set consisting of 29 attributes and containing various types of information such as dinucleotide frequencies, thermodynamical characteristics, statistical and topological factors. Very high performance was achieved and a novelty was the fact that they tested the model to a dataset derived from multiple organisms and other non-coding RNAs.

Two years later **microPred** [26] was an extension of miPred methodology. As in the miPred approach, the authors of microPred applied a SVM classifier to an extended dataset and they added 19 newly introduced features. They also applied some classic filtering methods to reduce the feature subset size. Moreover, the class imbalance problem, as described in Veropoulos et al. [27] was analyzed and imbalance learning experiments were also conducted. However, the basic limitation of the previous methods was the fact that they did not propose an advanced methodology to tune the parameters of the models.

In order to combine the improvements of different approaches, the **MiRFinder** software [28] introduced an alternative technique for genome wide miRNA identification. The model incorporated a SVM filtering step using 18 features and achieved lower performance than it was expected. In 2010, a hybrid methodology [29] which combines the efficient SVM classification with Genetic Algorithms optimization and feature selection was proposed. This method extracted a minimum prediction feature subset and it was integrated in an online tool named as ncRNA-class tool, which apart from predicting miRNAs, computes informative features applicable to other non-coding RNA classes. The idea of SVM classification with GA optimization was also adopted by the **miPred GA** methodology [30]. The problem of the better detection of miRNAs was handled in **MiRenSVM** deploying a SVM-based approach trained on several multi-loop features. Also the method studied the feature selection problem and applied statistical measures to extract features with high discriminative power. Recently, the **miRD** method [31], based on SVMs and a boosting method, focused on the feature selection strategies.

2.2.2. Probabilistic models

The ProMiR method [32] introduced a probabilistic co-learning model for miRNA gene identification. The authors analyzed the structure and the sequence of pre-miRNAs and a HMM capable of describing miRNA candidate was implemented. The method was the first to apply a probabilistic model and it included experimental confirmation of the results. The low sensitivity and the fair overall performance were the basic limitations. In the next years, extensive studies on miRNA structures and stem loops topology led to more sophisticated probabilistic models. Striking examples are the **HMMMiR** model [33] which integrated different algorithms for parameters estimation and the **CSHMM** method [34] which incorporated a context sensitive HMM model tested in both normal and Next Generation Sequencing (NGS) data. A Bayesian probabilistic model was applied in several cross species for effective miRNA prediction. The **BayesMirFind** method [35] applied a Naïve Bayes classifier on structural and sequential data from various species. The method by using a large and representative feature set reduced the number of false positives but it achieved lower performance compared to other techniques. One possible explanation for the low performance results is the strong hypothesis which is made by Naïve Bayesian approaches about the features independence.

2.2.3. Other techniques

The **miR-KDE** method [36] applied a relaxed variable kernel density estimator (RVKDE) and features collected from previous works to classify miRNA genes. The evaluation process shows that this method has advantages for miRNAs taxonomically distant to human. The **miRank** algorithm [37] introduced a technique suitable for cases when limited training examples are available. The algorithm was based on random walks and required very few known miRNAs for the learning phase. The experimental results showed the method was suitable for not well annotated genomes but the work did not include practical comparisons with existing methodologies. The problem of result interpretability was studied in the **G^{2}DE** method [38]. A novel kernel based classifier (Gaussian density estimator) was applied and the algorithm achieved performance comparable to the prevailing classification algorithms. Recently a classifier based on Random Forests was introduced in [39]. The method used positive and negative hairpins collected from previous works and 24 features that measure the compactness of the stem loop structure. The practical comparison showed performance comparable to Triplet-SVM and microPred in plant and virus independent datasets.

2.3. Methods that rely on high throughput techniques and NGS data

The present advances in high throughput techniques discovered new areas of research for the class of the small non-coding molecules. The huge amount of data produced by the NGS Technologies shed light to the way researchers analyze and understand the biological function of molecules such as the pi-RNA, the sno-RNA and the microRNAs.

Novel analysis techniques for small RNAs were proposed in [40]. A computational pipeline for studying the mechanism and function of small molecules using high throughput datasets was proposed. van Dongen et al. [41] using expression profiles developed an algorithm which can be applied to any ordered list of RNA or DNA sequences and reveal miRNA and siRNA binding sites. Moreover, the identification of mature miRNA coding regions is another opened-research area. Methods such as **MiRPara** [42] tried to integrate characteristics derived from pri-miRNA, pre-miRNA and mature miRNA to achieve more reliable results. Other functional web tools such as **DSAP** [43] provide multi-task services for the miRNA analysis and have the potential to accelerate the development of new NGS-based techniques. Especially for the prediction of microRNA genes using NGS data several prediction models have been developed. The **miRDeep** [44] was developed to discover active known and novel miRNAs from deep sequencing data extracted from Solid, Illumina and 454 platforms. A probabilistic model to score the compatibility of the position and frequency of sequenced RNA was used and the method reported hundreds of previously un-annotated miRNAs, four of which were validated by experimental techniques. The principles of **miRDeep** were used in **miRDeep2** [45] which reported hundreds of novel miRNAs. The previous methods influenced more specific and advanced methodologies such as **miRDeep-P** [46], a computational tool for the prediction of microRNA transcriptome in plants. The **miRanalyzer** [47] introduced an integrated tool for the analysis of NGS experiments. The tool, using annotated sequences in miRBase and matches in other transcribed sequences, introduced a prediction algorithm based on random forests. Finally, in 2010 **MiREna** [48], a search algorithm is proposed which was designed for the prediction of mature miRNAs and pre-miRNAs using five physical and combinatorial parameters.

3. microRNA target prediction

A large number of computational approaches have already been developed for miRNA target prediction and their characteristics and limitations have been reviewed in some previous works. Mendes et al. [10] discuss methods for miRNA gene finding and target identification both for animal and plants. Concerning the current computational techniques, they propose the usage of mRNA expression data and the incorporation of mRNA secondary structure in order to achieve better specificity. Maziere and Enright [49] address the limitations of the existing computational techniques and compare several approaches mainly in terms of web accessibility. They believe that the better understanding of the miRNA binding biology and the increase of the number of validated targets will accelerate the performance of the computational techniques. Similarly, in [50], features related to animal miRNA targeting are analyzed. In this study, the authors enumerated, presented and classified in terms of web accessibility existing tools which rely on a combination of seed matching, site accessibility, evolutionary conservation features and expression profiles characteristics. The fact that none of the existing tools has managed to incorporate successfully all the known features was also mentioned. In [51] the principles of microRNA target prediction based on empirical evidence was discussed and computational methods based on base pairing patterns, secondary structure, nucleotide composition of target sequences and evolutionary conservation were reviewed. Furthermore web-based methods which combine computational prediction with expression data were analyzed. The authors concluded that the trade-off between sensitivity and specificity has not been handled successfully by existing tools. Computational studies about miRNA gene finding, miRNA target prediction and regulation of miRNA genes are reviewed by Li et al. [11]. Specifically in this study, the existing tools were divided in two generations: those based on complementarity, thermodynamic characteristics or conservation and the machine learning-based ones.

In our review paper, in order to enumerate and describe the various methods of miRNA target prediction, we firstly classified them in two main categories: the filtering based approaches and the ML based ones. The earliest approaches were based on filtering, since only a small number of miRNA targets were known. The filtering approaches attempted to detect new targets, validate them experimentally and thus increase the pool of known targets. The basic idea of filtering was to produce a number of putative targets based either on complementarity or on thermodynamic characteristics. The earliest approaches were based on strict matching in specific 3′UTR positions which are called seeds. However, seed matching features are not sufficient for the effective prediction of miRNA targets because there exists other important factors such as the accessibility of the seed area which are crucial for a miRNA:target interaction. For this reason, a variety of non-seed features have been proposed in the literature. In addition, recent findings indicated miRNA binding in promoters, coding regions or 5′UTR regions and this necessitates the incorporation of such information to the existing methods. Using a combination of filtering steps and a variety of conservation, nucleotide composition and secondary structure criteria, filtering methods attempted to rank the putative targets from those of higher confidence to those of lower. As the number of known targets of the characteristics which were used as inputs for their prediction was growing, ML approaches emerged. These methods classify miRNA:mRNA interactions to those that represent true targets and those that are non-targets. Features of these miRNA:mRNA interactions including thermodynamical, structural, positional and other features are gathered and computed. Then a classifier is built to fit these features and the constructed model is capable of predicting true targets. In other words, in these methods the rules are conducted by the data. In the following, representative tools of both categories are described. In Table 2 the tools' category, their main advantages and disadvantages and their website are presented.

3.1. Filtering approaches

Stark et al. [52] proposed a method for the prediction of miRNA targets of *D. melanogaster*. Comparing *D. melanogaster* and *D. pseudoobscura* 3′ UTRs, they built a database of conserved 3′ UTRs, which they searched for sequences complementary to the seeds of miRNAs, allowing G:U mismatches. These sequences were then evaluated for their ability to form energetically favorable RNA:RNA duplexes with the miRNAs. In [53] by following a similar approach, the authors implemented a method called **miRanda**, which managed to predict 9 out of 10 previously validated targets of *D. melanogaster*. It requires complementarity for the whole miRNA, but gives a higher weight to seed complementarity, allowing G:U mismatches, inserts and deletes. The next step is the calculation of the free energy of the duplex. Conservation of miRNA-target pairs in *D. pseudoobscura* and *Anopheles gambiae* is finally used as a filtering step. Both of the above methods refer to the existence of multiple target sites in an mRNA as an additional filtering step to reduce the false positive rate.

TargetScan [54] is a filtering methodology that requires perfect seed complementarity. For each miRNA:target site duplex, its folding free energy is computed and then the UTRs are assigned a score taking into account multiple target sites. The conserved UTRs with the highest scores are assumed as targets. This method predicts miRNA targets conserved across multiple genomes such as human, mouse, rat, and pufferfish. When testing 15 target sites, 11 managed to be validated experimentally. An extension of **TargetScan** is **TargetScanS** [55]. In TargetScanS the perfect seed complementarity now concerns 6 nucleotides instead of 7 and UTRs without multiple target sites are not punished so strictly. For the conservation step, the pufferfish UTRs are not used, while dog and chicken UTRs are used. Although both TargetScan and TargetScanS manage to reduce their false positive rates, they tend to miss those true positive targets that do not have perfect seed complementarity or are not conserved across species.

Contrary to the previous methods which are based on the complementarity between miRNAs and their potential targets, **Diana-MicroT** [56] relies firstly on thermodynamic stability. It uses a 38 nucleotide window to search across the UTRs for potential binding sites. For every miRNA-target site pair, the minimum binding energy is computed using a dynamic programming algorithm that computes free energies for both Watson–Crick pairs and G:U mismatches. Conserved target sites in human and mouse are predicted without taking into account multiple target sites. The whole set of 7 target sites that were tested were also experimentally validated. Furthermore, in [57] the implementation of **RNAHybrid** is described. The method relies on thermodynamic stability and predicts multiple potential binding sites by finding the energetically most favorable hybridization sites of miRNAs in mRNAs. A statistical model is then built talking also into account conservation among *D. melanogaster*, *D. pseudoobscura* and *A. gambiae*. Targets of Drosophila are predicted in the 3′ UTRs and the coding sequence, as well.

The predicted targets of the above mentioned methods have been frequently used as an initial set of putative targets given as input to other filtering methods. **mirWIP** [58] ranked predicted targets of RNAHybrid based on seed matching, structural accessibility and thermodynamic characteristics. Moreover, **HOCTAR** [59] ranked predicted targets of miRanda, TargetScan and PicTar [60] based on expression correlation.

The problems of miRNA targeting out of the 3′UTR region as well as the validation of the results were cases of study for the miRWalk [61] methodology. The authors in order to capture interactions that occur not only in the 3′UTR implemented a filtering approach capable of identifying seeds in promoters, coding regions and 5′UTR of known genes. A scanning phase on the gene sequences searches for candidate binding sites and the extracted re-

Table 2
Most important computational methods for the miRNA target prediction which may be accessed through a web interface or are public available.

Method	Category	Advantages	Disadvantages	Website
miRanda	Filtering	High interpretability Uses features from various categories Not stringent seed matching Multiple target sites	Low prediction performance Conservation restrictions	http://www.microrna.org
TargetScan	Filtering	High interpretability Uses features from various categories Multiple target sites	Low prediction performance Conservation restrictions Stringent seed matching	http://www.targetscan.org/
Diana-MicroT	Filtering	High interpretability Thermodynamic stability Not stringent seed matching	Low prediction performance Conservation No multiple target sites	http://diana.cslab.ece.ntua.gr/microT/
TargetBoost	Machine learning	High prediction performance High interpretability	Does not incorporate knowledge from various sources	http://www.interagon.com/demo/
miTarget	Machine learning	High prediction performance Uses as inputs features from various categories Provides function evaluation of predicted targets	Only a simple filtering feature selection method is deployed Low interpretability	http://cbit.snu.ac.kr/~miTarget
MirTarget2	Machine learning	High prediction performance Uses as inputs features from various categories plus expression profiles data Stores results in a database	No feature selection procedure Low interpretability Missing values	http://mirdb.org/miRDB/
TargetSpy	Machine learning	Use a plethora of features as inputs High prediction performance	Absence of seed matching features Low interpretability Only a simple filtering feature selection	www.targetspy.org
Saito & Saetrom	Machine learning	High prediction performance Uses as inputs features from various categories plus expression profiles data Prediction models are attempted to be interpreted	No feature selection procedure Missing values	http://tare.medisin.ntnu.no/mirna_target
TargetMiner	Machine learning	Advanced negative set selection proposed Uses as inputs features from various categories plus expression profiles data	Low prediction performance Only a simple filtering feature selection method is deployed	http://www.isical.ac.in/~bioinfo_miu/
MultiMiTar	Machine learning	High prediction performance Uses as inputs features from various categories Includes and efficient feature selection procedure Handles the sensitivity–specificity trade-off	Small training dataset No web interface available	www.isical.ac.in/~bioinfo_miu/multimitar-download.htm
NBmiRTar	Hybrid	Mediocre prediction performance Uses as inputs features from various categories High interpretability	Only a multivariate filtering feature selection method is deployed Cannot use features with mutual information	http://www.biosino.org/~kanghu/mRTP/mRTP.html
microT-ANN	Hybrid	High prediction performance Uses as inputs features from various categories Negative training samples are computed efficiently using protein expression experimental results	Prone to overfitting Positive training samples are not selected properly No feature selection procedure is used No web interface available	http://microrna.gr/microT-ANN/

sults are combined with the results of eight well know tools for miRNA target prediction. The outcome that is stored in a relational database provides a more general overview of the miRNA targeting mechanism and gives us a holistic view of cellular interactions. Regarding the validation issue the authors developed a text mining approach which searches the abstracts of the PubMed articles and links the predictions with diseases, cellular pathways and OMIM disorders. The software encapsulates important properties and provides a fast and efficient tool for further comprehensive studies.

Similarly to the previous methodology the RNA22 [62] program tackles the problem of finding "non-canonical" miRNA targets and enables a variety of interactive functions incorporated to a user-friendly interface.

The main disadvantage of filtering methods is that in their attempt to reduce the false positive targets they use conservation and stringent seed matching criteria. As a result, they lose possible targets that may not be conserved or have seeds with mismatches. Filtering methods have, in general, low prediction performance, but high interpretability.

3.2. ML approaches

For the computational prediction of miRNA targets the ML techniques which have been applied include statistical techniques, Bayesian classifiers, Artificial Neural Networks, SVMs, Ensemble classifiers and a few advanced hybrid techniques which combine a meta-heuristic technique with a classifier.

The **MicroTar** [63] is a method which attempted to tackle the problem of predicting non evolutionary conserved targets. The tool was based on complementarity and thermodynamical data and introduced a pipeline which at the final step applies statistical analysis (EVD: Extreme Values Distributions) to extract candidate miRNA targets. The core of the algorithm relies on the miRNA:mRNA hybridization energy and the Vienna RNA package was used for the Minimum Free Energy (MFE) estimations. This methodology provided biological insight about the importance of the studied features in the prediction of miRNA targets and its parallel implementation enables large scale analysis. However, being a simple statistical method it provided mediocre predictive performance.

Gaidatzis et al. in [64] proposed a Bayesian model for predicting target genes. They studied the phylogenetic evolution of targets among species and their work included a functional pathway analysis using the KEGG database. This methodology was incorporated to the **MirZ** web server [65]. The server integrated miRNA expression profile data combined with target prediction results. A more advanced Bayesian approach was **GenMIR++** [66]. In this methodology a Bayesian model was developed which identified a network consisting of human mRNA targets and their miRNAs regulators. In order to validate their findings, they used Gene Ontology (GO) annotations and they further conducted experiments for the predicted let-7b targets. The paper except for the target prediction methodology established important knowledge regarding the tumor suppressor miRNAs. Such approaches are limited by the feature independence assumption which should hold for every Naïve Bayesian approach.

The **MTar** method [67] introduced an Artificial Neural Network architecture for predicting miRNA targets in the human transcriptome. A multi-layer perceptron was used to divide the target sites into three different categories and the classification included 16 positional, thermodynamic and structural attributes. By applying different thresholds for each target category they achieved high performance in terms of accuracy, specificity and sensitivity. Interesting is the way they generated the negative examples because they observed that deletions of targets positions produce large number of negative samples. Their most important problems were the difficulties in optimizing the neural network topology and parameters and the questing about the generalization abilities of Artificial Neural Networks.

The most important SVM approaches for the prediction of miRNA targets are miTarget [68], MirTarget2 [69] and TargetMiner [70]. The **miTarget** methodology [68] incorporated 41 structural, thermodynamic and position-based features. The authors attempted to include information taken from microarray experiments and novelty was the fact that they validated their prediction results using the GO. Finally, they studied the contribution of individual features and their analysis revealed the role of the thermodynamic features to the overall performance enhancement. The **MirTarget2** methodology [69] relied on the Linsley microarray transcriptional dataset to identify statistically significant features between downregulated and overexpressed mRNAs. The authors also presented the miRDB database which was designed to store the predicted targets and their functional annotation. The problem of selecting representative negative examples was studied extensively during the development of the **TargetMiner** methodology. This method proposed a flowchart for systematic identification of negative examples. The authors validated their findings by applying a variant of isotope labeling called pSILAC. Regarding the classification methodology they applied SVM with RBF kernel function on a set of 90 attributes. They also tackled the problem of feature selection by using the statistical F score. All these methodologies provided high predictive performance however they encountered some problems concerning classification models interpretability and parameter tuning.

MultiMiTar [71] attempted to tackle the limitations of existing SVM approaches for the prediction of miRNA targets. It is a wrapper methodology consisted of a multi-objective Simulated Annealing technique and an SVM classifier. This methodology is able to approximate the optimal feature subset which should be used as inputs for the classification model. Furthermore, the usage of multi objective optimization enables it to handle the sensitivity–specificity tradeoff and extract balanced classification models. Despite its theoretical superiority over the previous methodologies, it uses only a small training set of 187 positive and 57 negative examples which are not sufficient for the efficient training and testing of the extracted prediction models. Contrary to MultiMiTar, a more recent method [72] uses an equal number of positive and negative examples. It is a hybrid methodology, combining Genetic Algorithms with SVMs. This approach locates the optimal feature subset and achieves high classification performance, handling on the same time the sensitivity–specificity tradeoff. In specific, this method uses an advanced multi-objective fitness function to accomplish all the aforementioned goals. Moreover, it deploys an adaptive variation operator to improve the algorithm's convergence behavior.

In contrast to the conventional thermodynamic and complementarity classification attributes, the **TargetBoost** methodology [73] introduced the usage of weighted sequence motifs to characterize the binding sites between candidate miRNA genes and targets. The method relies on a boosted genetic programming algorithm and uses a grammar written in Backus Naur Form [74] to extract the target sites. Comparisons between this method and conventional filtering methodologies showed that it is capable of extracting significant knowledge regarding the miRNA:mRNA interactions. Genetic Programming is an extremely promising technique but it currently suffers from overfitting issues, slow convergence and the bloat effect. These problems stringed the performance of the overall methodology.

In [75], emphasis was given on the refinement of the target prediction results and presented a workflow based on ML methods. Using data from public available databases, they tested the Naïve Bayes classifier, Neural Networks, SVM and Decision trees on features derived from different categories. Their analysis showed that the SVM classifier achieved the best performance. Then they attempted to improve the classification performance by using the meta-algorithm called Adaboost.

Another ensemble methodology is **TargetSpy** [76], which focused on the efficient target prediction without the usage of seed matching attributes and conservation filtering. Thus, in order to include a representative feature set capable of predicting targets to a broad range of species, they collected a variety of features related to the miRNA:mRNA duplex characteristics. The core of their prediction methodology is the learning scheme called MultiBoost [77]. Regarding the importance of individual features they ranked the features using the ReliefF algorithm and they further extracted the optimal feature subset by applying correlation-based feature selection. Both ensemble methodologies are able to provide extremely high predictive performance, but provide prediction models with low interpretability.

3.3. Hybrid approaches

Except from the aforementioned basic categories of filtering and ML computational methodologies for the prediction of miRNA targets, lately some of the proposed methodologies include both filtering and ML steps. Thus, these methodologies are classified in a third category, which we named as hybrid approaches.

In [78] the mirTarget tool was used to deploy a weighted scheme that combines various conservation, sequential, thermodynamics and structural scores for every miRNA:mRNA pair. Then the predicted candidate miRNA targets were filtered out using the expression profiles of miRNAs and mRNAs.

The **NBmiRTar** program, proposed in [79], implemented a Naïve Bayes classifier to validate mRNA targets and artificially generated samples. The methodology combined Naïve Bayes algorithm features extracted from the sequences, the seed and out-seed regions and the duplex structures, and the complete pipeline contains additional filtering steps according to the miRanda software score. Contributions of the paper are considered the method for generating artificial negative example as well as the study of features with significant mutual information.

The most recent hybrid approach was proposed by Reczko et al. [80] in 2012. They introduced a novel tool, named **DIANA-microT-ANN**, which at first filters candidate target genes for each miRNA using some binding features of the 3′UTR region and then deploys an Artificial Neural Network classifier to combine a set of conservation, binding and structural accessibility features in order to predict the final set of the target genes. The positive and negative datasets were created using experimental evidence about the change of protein expression levels after the over-expression of miRNAs. However, the determination of positives using only this expression criterion is quite problematic as the change of the expression level of a specific protein may not be directly associated with a targeting relation between the examined miRNA and the mRNA which is expressed to this protein.

4. Computational analysis of miRNA transcription procedure

Existing methodologies that have already been applied for the prediction of miRNA genes and for the prediction of their target genes were presented in previous sections. Another emerging research direction is the study of the mechanisms which control the regulation of the miRNA genes. Specifically, not only there exist mechanisms that involve miRNAs in the regulation of their target genes but there also exist mechanisms that involve miRNA and other genes in the regulation of the miRNA genes themselves.

Experimental studies have already conformed that most of the miRNA genes are transcribed by RNA polymerase II [81]. The promoters of each miRNA gene and its binding sites should be predicted in order to gain insight of such mechanisms and try to understand them. Recently, a variety of computational methodologies have been developed and applied to fulfill these tasks.

Zhou et al. [82] developed a classification approach, using SVM classifiers and sequence motifs as features, which is able to distinguish genes transcribed by RNA polymerase II and genes transcribed by RNA polymerase III. Their experimental results indicated that most of the examined microRNAs in four species (*C. elegans*, *Homo sapiens*, *A. thaliana*, and *O. sativa*) are transcribed by RNA polymerase II which is something that was expected due to existing beliefs and experimental evidence. However, a small number of microRNAs were found to be transcribed by RNA polymerase III. These results should be validated experimentally in order to ascertain if they are just an error of the classifier or true biological knowledge. Moreover, in [82] there has been developed a decision tree variation, named common query voting (**CoVote**), for the prediction of putative promoters of intergenic microRNAs. Specifically, for each microRNA genes the algorithm examines using a sliding window approach the upstream sequence and using decision trees with sequence motif features determines whether a sequence is a putative promoter or not. The prediction accuracy of CoVote was over 84% in all examined organisms.

More recently, in [83] a study in which ChIP-seq experiments with four transcriptional factors in human and mouse cells was conducted. The authors combined information of proximity to annotated mature microRNA sequences, available embryotic cell transcription data and conservation between species and they produced an ad hoc score for every putative transcription start site. Later, Corcoran et al. [84], developed a new tool (Core Promoter Prediction Program – **CCCP**) to identify putative core transcriptional start sites of the microRNA genes using Chip–Chip data. This method applies a SVM classifier with linear kernel function. The features which were used are appearance of DNA binding profiles, appearance of n-mers (with n having values of 3 and 4) and GC content. Some feature subsets were examined and the experimental results showed that the combination of the n-mers features with the GC content produces the highest classification results. In comparison with the Marson et al. approach, experimental evidence showed that CCCP algorithm is more accurate in predicting microRNA transcriptional start sites. This could easily been explained by the higher quality data that they used (Chip-seq suffers from the high presence of noise in microarray experiments) and the most advance classification algorithm that was deployed. The latest approach for the computational identification of miRNA transcriptional start sites is the one proposed in 2011 in [85]. Similarly to [84], it is mainly an SVM classifier, but it uses a high quality human dataset extracted from experimental evidence such as gene expression (CAGE) tags, transcriptional start sites Seq libraries and H3K4me3 chromatin signature derived from high-throughput sequencing analysis.

In 2010, Megraw and Hatzigeorgiou [5] studied the computational identification of transcriptional factors binding sites for microRNA promoters in plants. 63 experimentally verified microRNA transcriptional starting sites in *Arabidopsis* were studied. The areas of 800 nucleotides upstream these miRNA transcriptional starting sites was searched for known transcriptional factor binding sites. This search was conducted using a novel approach that deploys Positional Weight Matrices. A small number of factors (AtMYC2, ARF, SORLREP3 and LFY) were found significantly more frequently in miRNA promoters than in protein coding genes promoters.

5. miRNA mediated gene regulatory networks

Genes and their products are governed by regulation relationships among them. For years, researchers have focused on the accurate construction of gene regulatory networks with protein-coding genes being nodes for these networks. It is a common belief that these networks can provide useful information regarding the cell's function and the genes transcription mechanisms. Many methodologies have been proposed in the literature, such as dynamic models, network topology models and network control logic models, and all of them are based on high-throughput or low-throughput experimental data [86]. The discovery of miRNA genes and their transcription role, in addition to the early findings about their regulation procedures, has inspired researchers to endeavor the construction of gene regulatory network with nodes that represent both protein-coding genes and miRNA genes.

In 2010, Herranz and Cohen [6] reviewed the first attempts to incorporate miRNA genes in small-scaled specific gene regulatory networks. They provided extremely useful conclusions about the participation of miRNA genes in specific biological processes such as glucose homeostasis, photoreceptor differentiation, neuronal differentiation and many more. Specifically they emphasized on the ability of miRNAs to buffer the consequences of environmental noise providing robustness to environmental changes.

One of the first interesting attempts to associate groups of miRNA genes with mRNAs which constitute the most important part of gene regulatory networks was conducted in [87]. This approach, aimed at locating small scale miRNA:mRNA modules and examining their functionality. To achieve this scope they used computationally predicted targets for the examined miRNAs and expression profiles information. For the analysis of the data and the discovery of miRNA:mRNA modules they developed a novel co-evolutionary methodology that evolves on parallel populations of miRNAs subsets and mRNA subsets, evaluates their combinations through a novel evaluation function, and differentiate candidate solutions using a probabilistic operator.

In [88], a general workflow for the construction of a miRNA-mediated gene regulatory network in plants was proposed. Their workflow's main steps are the collection of miRNA promoters, the discovery of transcription factors–miRNA relationships, the computational prediction of miRNAs targets, the determination of genes expression profiles and the examination of self-regulation procedures for miRNAs. For all these steps, they propose alternative computational and experimental methodologies. As a case study, they applied this workflow for the determination of miR399-mediated gene regulatory networks for the plants *Arabidopsis* and rice.

In [89], the authors made a first effort to build a large scale gene regulatory network consisting of transcriptional factors, miRNA and protein coding genes. This study focused on the *C. elegans* organism but their results were verified in human and mouse organisms. Specifically, they used ChIP-Seq data for the determination of interactions between transcriptional factors and genes, RNA-seq profiles to classify transcription factors to positive and negative regulators, conservation information to predict target genes of miRNAs and at finally they incorporated protein–protein interactions information and potential intra-regulations among miRNAs. Despite the fact that some incorporation steps could be optimized by using modern methodologies concerning the more accurate determination of target genes for the miRNAs and of protein–protein interactions, this approach was among the first studies that offered a large-scale integration framework for the construction of gene regulatory networks with miRNA genes.

6. Discussion

In this section, after presenting the most important computational techniques for the prediction of miRNA genes/targets and for analyzing their regulatory role, some limitations and possible obstacles are discussed. Furthermore, directions for the future research agenda are proposed.

The miRNA data analysis includes algorithmic classifiers and predictors in most of the aforementioned analysis steps such as the prediction of miRNAs and the prediction of their targets. Among the computational methods which were presented in Tables 1 and 2, a clear answer to the dominant methodology does not exist. Despite the high performance of some of the existing methodologies, only a few of them are able to provide biological insight about the way that the input features are combined to extract the final prediction. Specifically, modern ML techniques such as SVM and Random Forest achieve high prediction performance but their extracted prediction models cannot be interpreted to extract biological inferences. Furthermore, simple statistical and filtering techniques are easier to be interpreted but produce mediocre prediction results. To handle this tradeoff, some recent methodologies which are able to interpret complex ML techniques should be used [90].

Furthermore, even the performance of the best methodologies among the ones presented in previous sections can be enhanced. The problems that should be solved to achieve this goal are the negative set selection, the class imbalance problem and the feature selection problem. When a classifier is to be trained, it is very crucial to use high quality training and test sets. The two most important classification problems examined in this paper are the classification of hairpins in pseudo ones and miRNAs and the classification of miRNA:mRNA pairs in targeting and non-targeting ones. The selection of positive examples for these classification results is straightforward by using information included in public databases such as miRBase and TarBase. Furthermore, the selection of pseudo hairpins is also simple as they can be downloaded from RefSeq genes. In opposite, the selection of pairs of miRNAs and mRNAs which do not share a targeting relationship is very difficult. The optimal methodology proposed so far, includes the usage of artificially generated miRNAs which are used as inputs in existing target prediction software to predict their targets. These miRNA:mRNA pairs are used as negative examples because they share similar properties to positive ones but are in fact negative as their miRNAs do not exist. This methodology has been used in [79] with great results. The methodology used in TargetMiner systematized the problem of identifying negative examples. However, their methodology for acquiring negative examples is mainly computational and thus the retrieved negative samples still include some false negatives. Thus, the problem of modeling negative samples with similar characteristics to the true non-target genes remains and interesting topic. We expect that the increase of the known experimentally verified negative examples will lead to more sophisticated modeling methods and will enhance the classification performance. Recent findings indicated that miRNAs can also control the expression of their target genes by base-pairing within the Promoter, 5′-UTR and CDS regions. Information about matching in the promoters, 5′UTR or CDS regions will enhance the prediction performance and on that direction databases such as miRWalk [61] and RNA22 [62] encapsulate important properties which could be used.

Another important limitation of existing methodologies is the class imbalance problem. Specifically, in nature pseudo miRNAs are much more than real miRNAs and miRNA:mRNA duplexes which do not share a targeting property are much more than miRNA:mRNA target duplexes. Thus, the problem that emerges is to find the optimal positive to negative samples that should be used to obtain the highest performance. The most commonly used approach that has been used in the miRNA analysis problems is the one that starts from 1:1 positive to negative samples rate and experimentally checks if increasing this rate up to 1:10 can increase the method's performance. The determination of the optimal negative to positive rate using the previous methodology is conducted through experimentation in the training set and may thus lead to overfitting. Moreover, the application of simple oversampling or under sampling methods to handle the class imbalance problem may result in extracting very specific classification boundaries and thus lead to overfitting. For these reasons, modern alternative class imbalance methods, such as SMOTE, should be used.

All the miRNA analysis steps which included an algorithmic predictor step are highly sensitive to the inputs which should be used in these predictors. The candidate inputs may range from simple sequential ones, to complex structural, thermodynamical and functional ones. Using low informative inputs or inputs with identical information may deteriorate the algorithms' performance and make the results interpretation a very difficult task. This is the reason why a feature selection algorithm should be applied to select the optimal feature subset which should be used as inputs for every prediction problem. In the miRNA analysis tasks, most of the times the researchers select their features empirically; or by using simple filtering techniques which take into account the feature dependencies and their interaction with the prediction algorithm. A few wrapper approaches have been proposed and most of them deploy a Meta-heuristic algorithm such as a Genetic Algorithm to estimate the optimal feature subset which should be used by the examined classifier. However these approaches are time consuming and may lead to overfitting. To overcome these difficulties modern embedded methods which take advantage of internal classifiers characteristics, should be used.

Another important problem that arises from the analysis of miRNA data is the comparison between existing methodologies. Most of them provide experimental results from different datasets and using different evaluation metrics. This makes it difficult to compare these methods in practice and restrict us on a practical comparison based on their quality characteristics. Thus, the production of public databases, which will contain benchmark datasets for the miRNA analysis tasks, is nowadays essential.

Another restrictive factor for the comparison of existing methodologies is the absence of web tools for many of the existing methodologies. Moreover, some of the existing web tools enable users only to predict at once only one miRNA or one target or one transcription factor. This fact makes their usage practically impossible. Hence, it is highly recommended for novel methodologies to be incorporated in user friendly web interfaces and be provided through web services in order to enable their extensive usage.

As already mentioned in this manuscript the emphasis that has been given in the development of computational techniques for the prediction of miRNA genes and their targets, enabled researchers lately to use their results and to try to solve most advanced issues such as uncovering the transcription mechanism of miRNAs and building miRNA mediated gene regulatory networks. However, up to our knowledge there does not exist an integrative approach to state-of-the-art algorithms from all the steps of the miRNA analysis procedure. This could be a very interesting future direction in order to gain more extensive insight on the mechanisms which govern the function of miRNAs and their role in the underlying cellular procedures.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

[1] Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol 2012;9.
[2] Lai EC. MicroRNAs: runts of the genome assert themselves. Curr Biol 2003;13:925–36.
[3] Zhang B, Pan X, Wang Q, et al. Computational identification of microRNAs and their targets. Comput Biol Chem 2006;30:395–407.
[4] Yousef M, Showe L, Showe M. A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. FEBS J 2009;276:2150–6.
[5] Megraw M, Hatzigeorgiou A. MicroRNA promoter analysis. Methods Mol Biol 2010;592:149–61.
[6] Herranz H, Cohen S. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev 2010;24:1339–44.
[7] Grad Y, Aach J, Hayes GD, et al. Computational and experimental identification of *C. elegans* microRNAs. Mol Cell 2003;11:1253–63.
[8] Chen PY, Manninga H, Slanchev K, et al. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. Genes Dev 2005;19:1288–93.
[9] Lindow M, Gorodkin J. Principles and limitations of computational microRNA gene and target finding. DNA Cell Biol 2007;26:339–51.
[10] Mendes ND, Freitas AT, Sagot MF. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 2009;37:2419–33.
[11] Li L, Xu J, Yang D, et al. Computational approaches for microRNA studies: a review. Mamm Genome 2010;21:1–12.
[12] Bentwich I, Avniel A, Karov Y, et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 2005;37:766–70.
[13] Lim LP, Lau NC, Weinstein EG, et al. The microRNAs of *Caenorhabditis elegans*. Genes Dev 2003;17:991–1008.
[14] Lai EC, Tomancak P, Williams RW, et al. Computational identification of *Drosophila* microRNA genes. Genome Biol 2003;4:R42.
[15] Adai A, Johnson C, Mlotshwa S, et al. Computational prediction of miRNAs in *Arabidopsis thaliana*. Genome Res 2005;15:78–91.
[16] Dezulian T, Remmert M, Palatnik JF, et al. Identification of plant microRNA homologs. Bioinformatics 2006;22:359–60.
[17] Wang X, Zhang J, Li F, et al. MicroRNA identification based on sequence and structure alignment. Bioinformatics 2005;21:3610–4.
[18] Jones-Rhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 2004;14:787–99.
[19] Berezikov E, Thuemmler F, van Laake LW, et al. Diversity of microRNAs in human and chimpanzee brain. Nat Genet 2006;38:1375–7.
[20] Cullen BR. Viruses and microRNAs. Nat Genet 38(Suppl.):S25–30.
[21] Xue C, Li F, He T, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 2005;6:310.
[22] Sewer A, Paul N, Landgraf P, et al. Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 2005;6:267.
[23] Szafranski K, Megraw M, Reczko M, Hatzigeorgiou AG. Support vector machines for predicting microRNA hairpins. Proc Biocomp 2006:270–6.
[24] Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics 2006;22:197–202.
[25] Ng KLS, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 2007;23:1321–30.
[26] Batuwita R, Palade V. MicroPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 2009;25:989–95.
[27] Veropoulos K, Cristianini N, Campbell C. The application of support vector machines to medical decision support: a case study. ACAI 1999.
[28] Huang TH, Fan B, Rothschild MF, et al. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics 2007;8:341.
[29] Theofilatos K, Kleftogiannis D, Rapsomaniki M, et al. A novel pre-miRNA classification approach for the prediction of microRNA genes. In Proceedings of the 10th IEEE international conference on information technology and applications in biomedicine (ITAB); 2010.
[30] Xuan P, Guo M, Liu X, et al. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics 2011;27:1368–76.
[31] Zhang Y, Yang Y, Zhang H, et al. Prediction of novel pre-microRNAs with high accuracy through boosting and SVM. Bioinformatics 2011;27:1436–7.
[32] Nam JM, Shin KR, Han J, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res 2005;33:3570–81.
[33] Agarwal S, Vaz C, Bhattacharya A, et al. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinformatics 2010;11(Suppl. 1):S29.
[34] Kadri S, Hinman V, Benos PV. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinformatics 2009;10(Suppl. 1):S35.

[35] Yousef M, Nebozhyn M, Shatkay H, et al. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics 2006;22:1325–34.

[36] Chang DTH, Wang CC, Chen JW. Using a kernel density estimation based classifier to predict species-specific microRNA precursors. BMC Bioinformatics 2008;9(Suppl. 12):S2.

[37] Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics 2008;24:i50–8.

[38] Hsieh CH, Chang DTH, Hsueh CH, et al. Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm. BMC Bioinformatics 2010;11(Suppl. 1):S52.

[39] Xiao J, Tang X, Li Y, et al. Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure. BMC Bioinformatics 2011;12:165.

[40] Olson AJ, Brennecke J, Aravin AA, et al. Analysis of large-scale sequencing of small RNAs. Pac Symp Biocomput 2008:126–36.

[41] van Dongen S, Abreu-Goodger C, Enright AJ. Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods 2008;5:1023–5.

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

[43] Huang PJ, Liu YC, Lee CC, et al. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res 2010;38:W385–91.

[44] Friedlander MR, Chen W, Adamidi C, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 2008;26:407–15.

[45] Friedländer MR, Mackowiak SD, Li N, et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2011.

[46] Yang X, Li L. MiRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics 2011;27:2614–5.

[47] Hackenberg M, Sturm M, Langenberger D, et al. MiRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2009;37:68–76.

[48] Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics 2010;26:2226–34.

[49] Maziere P, Enright AJ. Prediction of microRNA targets. Drug Discov Today 2007;12:452–8.

[50] Saito T, Saetrom P. MicroRNAs targeting and target prediction. New Biotechnol 2010;27(3):243–9.

[51] Min H, Yoon S. Got target?: computational methods for microRNA target prediction and their extension. Exp Mol Med 2010;42(4):233–44.

[52] Stark A, Brennecke J, Russell RB, et al. Identification of *Drosophila* MicroRNA targets. PLoS Biol 2003;1:E60.

[53] Enright AJ, John B, Gaul U, et al. MicroRNA targets in *Drosophila*. Genome Biol 2003;5:R1.

[54] Lewis BP, Shih I, Jones-Rhoades MW, et al. Prediction of mammalian MicroRNA targets. Cell 2003;115(7):787–98.

[55] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005;120(1):15–20.

[56] Kiriakidou M, Nelson PT, Kouranov A, et al. A combined computational–experimental approach predicts human microRNA targets. Genes Dev 2004;18(10):1165–78.

[57] Rehmsmeier M, Steffen P, Hochsmann M, et al. Fast and effective prediction of microRNA/target duplexes. RNA 2004;10(10):1507–17.

[58] Hammell M, Long D, Zhang L, et al. MirWIP: microRNA target prediction based on miRNP enriched transcripts. Nat Methods 2008;5(9):813–9.

[59] Gennarino VA, Sardiello M, Avellino R, et al. MicroRNA target prediction by expression analysis of host genes. Genome Res 2009;19(3):490–1.

[60] Krek A, GrÃin D, Poy MN, et al. Combinatorial microRNA target predictions. Nat Genet 2005;37(5):495–500.

[61] Dweep H, Sticht C, Pandey P, Gretz N. MiRWalk – database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform 2011;44:839–47.

[62] Loher Phillipe, Rigoutsos Isidore. Interactive exploration of RNA22 microRNA target predictions. Bioinform Appl Note 2012;28(24):3322–3.

[63] Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioinformatics 2006;7:S20.

[64] Gaidatzis D, van Nimwegen E, Hausser J, et al. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics 2007;8:69.

[65] Hausser J, Berninger P, Rodak C, et al. MirZ: an integrated microRNA expression atlas and target prediction resource. Nucleic Acids Res 2009;37:W266–72.

[66] Huang JC, Babak T, Corson TW, et al. Using expression profiling data to identify human microRNA targets. Nat Methods 2007;4:1045–9.

[67] Chandra V, Girijadevi R, Nair AS, et al. MTar: a computational microRNA target prediction architecture for human transcriptome. BMC Bioinformatics 2010;11:S2.

[68] Kim SK, Nam JW, Rhee JK, et al. MiTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics 2006;7:411.

[69] Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics 2008;24:325–32.

[70] Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics 2009;25:2625–31.

[71] Mitra R, Bandyopadhyay S. MultiMiTar: a novel multi objective optimization based miRNA-target prediction method. PLoS One 2011;6(9):e24583.

[72] Korfiati A, Kleftogiannis D, Theofilatos K, et al. Predicting human miRNA target genes using a novel evolutionary methodology. Artif Intell: Theor Appl Lect Notes Comput Sci 2012;7297(2012):291–8.

[73] Saetrom O, Snove O, Saetrom P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA 2005;11:995–1003.

[74] Knuth DE. Backus normal form vs. Backus Naur form. Commun ACM 1964;7:735–6.

[75] Yan X, Chao T, Tu K, et al. Improving the prediction of human microRNA target genes by using ensemble algorithm. FEBS Lett 2007;581:1587–93.

[76] Sturm M, Hackenberg M, Langenberger D, et al. TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics 2010;11:292.

[77] Webb GI. MultiBoosting: a technique for combining boosting and wagging. Mach Learn 2000;40:159–96.

[78] Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res 2006;34(5):1646–52.

[79] Yousef M, Jung S, Kossenkov AV, et al. Naïve Bayes for microRNA target predictions-machine learning for microRNA targets. Bioinformatics 2007;23(22):2987–92.

[80] Reszko M, Maragkakis M, Alexiou P, et al. Accurate miRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. Front Genet 2012;2:103.

[81] Lee Y, Kim M, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J 2004;23:4051–60.

[82] Zhou X, Ruan J, et al. Characterization and identification of micro-RNA core promoters in four model species. PLoS Comput Biol 2007;3(3):412–23.

[83] Marson A, Levine S, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryotic stem cells. Cell 2008;134(3):521–33.

[84] Corcoran D, Pandit K, et al. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PLoS One 2009;4(4):e5279.

[85] Chien C, Sun Y, Chang W, et al. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. Nucleic Acids Res 2011;39(21):9345–56.

[86] Schlitt T, Brazma A. Current approaches to gene regulatory network modeling. BMC Bioinformatics 2007;8(Suppl. 6):s9.

[87] Joung J, Hwang K, et al. Discovery of microRNA–mRNA modules via population-based probabilistic learning. Bioinformatics 2007;23(9):1141–7.

[88] Meng Y, Shao C, Chen M. Towards microRNA-mediated gene regulatory networks in plants. Brief Bioinform 2011;12(6):645–59.

[89] Cheng C, Yan K, et al. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol 2011;7(11):e1002190.

[90] Papadimitriou S, Terzidis K. Efficient and interpretable fuzzy classifiers from data with support vector learning. Intell Data Anal 2005;9:527–50.