Chapter 2

Computational Identification of miRNAs Involved in Cancer

Anastasis Oulas, Nestoras Karathanasis, and Panayota Poirazi

Abstract

Changes in the structure and/or the expression of protein-coding genes were thought to be the major cause of cancer for many decades. However, the recent discovery of non-coding RNA (ncRNA) transcripts suggests that the molecular biology of cancer is far more complex. MicroRNAs (miRNAs) are key players of the family of ncRNAs and they have been under extensive investigation because of their involvement in carcinogenesis, often taking up roles of tumor suppressors or oncogenes. Owing to the slow nature of experimental identification of miRNA genes, computational procedures have been applied as a valuable complement to cloning. Numerous computational tools, implemented to recognize the characteristic features of miRNA biogenesis, have resulted in the prediction of multiple novel miRNA genes. Computational approaches provide valuable clues as to which are the dominant features that characterize these regulatory units and furthermore act by narrowing down the search space making experimental verification faster and significantly cheaper. Moreover, in combination with large-scale, high-throughput methods, such as deep sequencing and tilling arrays, computational methods have aided in the discovery of putative molecular signatures of miRNA deregulation in human tumors. This chapter focuses on existing computational methods for identifying miRNA genes, provides an overview of the methodology undertaken by these tools, and underlies their contribution toward unraveling the role of miRNAs in cancer.

Key words: MicroRNAs, Gene prediction, Software tools comparison, Cancer

1. Introduction

Current estimates show that while over 30% of vertebrate genomes are transcribed (1), only 1% represents protein-coding genes; the rest are believed to be various types of non-coding RNA (ncRNA) genes. Currently, only ~700 human microRNA (miRNA) genes exist in the miRNA registry,¹ and it is anticipated that there may be thousands more. The role of these molecules in cancer has

¹miRBase, release 13.0.
lately received a great deal of the scientific community's attention. Specifically, recent findings indicate that alterations in the expression of several miRNAs are often present in human cancers, suggesting potential roles of miRNAs in carcinogenic processes. For example, the expression levels of \( \text{let-7} \) \(^2\), \( \text{miR-15a/miR-16-1} \) cluster \(^3\), and neighboring \( \text{miR-143/miR-145} \) \(^4\) are found to be reduced in some malignancies, suggesting their potential role as tumor suppressors. In contrast, some other miRNAs such as the \( \text{miR-17-92} \) cluster \(^5–7\) and \( \text{miR-155/BIC} \) \(^8\) are overexpressed in various cancers, suggesting a possible oncogenic role. Furthermore, some miRNAs with altered expression levels appear to be associated with certain genetic alterations, such as deletion, amplification, and mutation. Regions that are prone to such genetic alterations are commonly referred to as cancer-associated genomic regions (CAGRs) and fragile sites (FRA) \(^9\). MiRNA genes located within or in close proximity to these regions have been suggested to be associated with chromosomal events leading to carcinogenesis, as graphically illustrated in Fig. 1.

The large amount of unexplored non-coding regions in the human genome combined with the increasing importance of miRNAs in cancer highlights the need for fast, flexible, and reliable miRNA identification methods. Toward this goal, a number of different computational methods have been used to identify

![Fig. 1. MiRNAs as cancer players. Computational prediction initiates the search for putative miRNAs that play a role in tumorigenesis. Some of these proposed mechanisms are experimentally proven, like the deletion of \( \text{miR-15a/miR-16a} \) cluster in B-CLL \((3, 49)\), the c-myc overexpression by the reposition near a putative miR promoter \((9)\), or \( \text{miR143/miR-145} \) cluster downregulation in colon cancers \((4)\). Figure adopted with permission from Calin et al. \((9)\).](image-url)
miRNA genes. Early studies focused on scanning for hairpin structures conserved between closely related species such as *Caenorhabditis elegans* and *Caenorhabditis briggsae* (10, 11), or using homology between known miRNAs and other regions in aligned genomes like human and mouse (12). Other approaches relied on conserved regions of synteny – conserved clustering of miRNAs in closely related genomes – to predict novel miRNAs (12). Subsequent computational studies utilized profile-based detection (13) as well as secondary structure alignment (14) of miRNAs using sequence conservation across multiple, highly divergent organisms (i.e., mouse and fugu).

The main drawback of the abovementioned tools is that they undertake a pipeline approach by applying stringent cut-offs and eliminating candidate miRNAs as the pipeline proceeds (10, 11). This results in the loss of numerous true miRNAs along the line. The use of homology by some tools (12–14) to detect novel miRNAs based on their similarity to previously identified miRNAs is another drawback. These methods obviously fall short when scanning distantly related sequences or when novel miRNAs lack detectable homologs.

The next generation of computational tools relied on more sophisticated machine learning algorithms such as support vector machine tools (SVMs) capable of taking into account multiple biological features such as free energy of the hairpin structure, paired bases, loop length, and stem conservation to predict novel miRNAs (15–17). Two very effective computational studies utilized hidden Markov models (HMMs) and a Bayesian classifier (18, 19) to simultaneously consider sequence and structure features at the nucleotide level for predicting miRNA genes. These studies, however, did not integrate conservation information in their algorithms, an important feature of the majority of miRNA genes. More recently, two computational tools miRRim (20) and SSCprofiler (21) also employing HMMs proved to be very effective, achieving high performance on identifying miRNAs in the human genome.

With the advent of large-scale, high-throughput methods such as tiling arrays or deep sequencing, the identification of novel miRNA genes is taking a different turn (22–24). These methods are exceptionally useful as they produce large datasets that offer a relatively accurate expression map for small RNAs in the genome. However, as large-scale expression data are usually limited by the specific tissue and developmental stage of their samples, only the coupling of such data to computational tools (as done in two recent studies (20, 21)) can facilitate rapid and precise detection of novel miRNAs while at the same time giving greater credence to computational predictions.
In the following paragraphs, we overview the representative examples of miRNA gene prediction tools and highlight their most important characteristics. The tools are organized according to the biological features they implement.

2. Tool Comparison

2.1. Sequence-Based Prediction

The initial computational tools for miRNA identification were based on sequence conservation with already cloned miRNAs. For instance, Quintana et al. (25) predicted 34 novel miRNAs using tissue-specific cloning. Almost all of these miRNAs were conserved in the human genome and frequently in non-mammalian vertebrate genomes such as pufferfish. One interesting observation was that certain miRNAs showed increased expression in heart, liver, or brain when compared with the entire miRNA population known at the time, proposing a role of these miRNAs in tissue specification or cell lineage decision.

2.2. Sequence, Structure, and Closely Related Species Conservation

Early computational methods for miRNA gene prediction relied on rules derived from sequence and structural features of miRNA precursors as well as their degree of conservation across species. The use of conservation was usually limited to pairwise comparison of closely related species. MiRscan (11) and MiRseeker (10) are two representative approaches for this category of tools. The MiRscan (11) tool implements a probabilistic method that uses known miRNAs as a training set in order to derive new miRNAs based on their degree of similarity. Specifically, the tool was developed as follows:

1. A total of 36,000 conserved sequences between C. elegans and C. briggsae (WU-BLAST cut-off $E < 1.8$) were scanned using RNAfold in search for hairpin structures.

2. Fifty of them, which have previously been reported as real miRNAs, were used as the training set in order to derive statistical information regarding a set of characteristic features (shown in Fig. 2).

3. Using the trained algorithm, all 36,000 sequences were evaluated with respect to their potential as miRNA precursors. Identification of new miRNA precursors by MiRscan is achieved via the use of a sliding window of 21 nt that is shifted across each hairpin candidate searching for the presence of specific miRNA features. These features include:

- Complementary base pairing;
- High degree of 5’ and 3’ conservation;
- Certain degree of bulge symmetry;
- Specific distance from the loop;
- Initial pentamer properties.
MiRseeker (10) is another tool belonging to the same category. It has been used for predicting miRNA genes in two Drosophila species, namely *Drosophila melanogaster* and *Drosophila pseudoobscura*, via the use of the following three-step pipeline approach:

1. The two genomes are aligned to find all conserved regions.
2. MiRseeker, which includes MFOLD, is used to identify miRNA genes. A very important feature of MiRseeker which enhances its prediction power is the consideration of pattern divergence. As per the authors’ report, there is less selective pressure and hence less conservation in the loop sequences.
3. The method is evaluated according to its ability to assign high scores to 24 known Drosophila miRNAs.

Another approach that has been used to predict novel human miRNAs is known as phylogenetic shadowing (26, 27). This approach uses multiple sequence comparisons of closely related species and allows the identification of conserved regions.
The method is based on observing conserved islands of a specific length, which are considered likely to be miRNA genes, in an otherwise unconserved region (26). An example of phylogenetic shadowing application to the identification of miRNA genes is found in (27), using the following steps:

1. Sequences from more than 100 miRNA regions in ten different primates are compared in order to infer a characteristic profile:
   - Variation in the loop sequences;
   - Conservation in stem of hairpins;
   - Significant decrease in conservation of sequence flanking the hairpins.

2. This profile is used to identify new miRNAs in pairwise alignments of more divergent species such as human and mouse or human and rat.

3. Additional filtering is performed according to the folding energy of candidate sequences.

A total of 976 potential human miRNAs have been identified using this method. This set contains over 80% of all known human miRNAs in version 3.1 of the miRNA registry (http://microrna.sanger.ac.uk/sequences/). Northern blot analyses combined with database searches reach a conservative estimate of 200–300 verified novel human miRNAs, a twofold increase over previous studies (28). Strong conservation over all species is only evident for two well-known miRNA genes.

Another approach is the use of full genome sequence alignment (12, 29, 30). The motivation being that human and mouse miRNAs should reside in conserved regions of synteny. A representative study using full genome alignments is found in (12), whereby:

1. The BLAT comparison tool (31) is used to compare the entire set of human and mouse precursor and mature miRNAs in the miRNA registry, version 2.2 (http://microrna.sanger.ac.uk/sequences/).

2. The results are further filtered using secondary structure prediction tools, like MFOLD and other criteria (such as G:U base pairing).

3. Characteristic features of some miRNAs are used to identify miRNA gene clusters and display conservation in the location of the clusters in comparison with other neighboring genes.

The findings of this work included the prediction of 80 new putative miRNAs genes.

The computational approaches described so far utilize information regarding closely related species and homology searches.
using sequences of already cloned miRNAs. These methods proved to be successful, leading to the prediction of multiple new miRNA genes. However, as they were bound by species similarity they eventually reached a prediction limit. To overcome this obstacle, researchers turned their attention to the prediction of miRNAs that do not show conservation to other known miRNAs and are not highly conserved across closely related species.

**2.4. A More General Model for miRNA Gene Prediction**

miRAlign (14) is one of the first tools that detects new miRNAs based on both sequence and structure alignments without implementing conservation features. The main characteristics that differentiate this tool from the existing homolog search methods are:

1. By applying a relatively loose conservation of the mature miRNA sequence, it has the ability to find distant homologs.
2. It considers more structural properties by introducing a structure alignment strategy that can use each single miRNA as a query for genomic searches.

The tool has been shown to perform better in comparison with other tools such as BLAST or ERPIN (32), and its main advantage is the prediction of more distant miRNA homologs or orthologs.

**2.5. The Next Generation of Tools**

2.5.1. HMM Tools

2.5.1.1. ProMir — Nam et al. (18)

1. A highly specific probabilistic co-learning method was hand-crafted, based on the paired HMM.
2. This method combines both sequence and structural characteristics of miRNA genes in a probabilistic framework and simultaneously decides if a miRNA gene and the mature miRNA are present by detecting the signals for the site cleaved by Drosha.
3. miRNA gene candidates are finally filtered using conservation across multiple divergent species.

Most recently, two freely available prediction tools (SSCprofiler (21) and miRRim (20)) have been shown to predict miRNA genes with high accuracy.

2.5.1.2. SSCprofiler Oulas et al. (21)

1. SSCprofiler utilizes profile HMM trained to recognize key biological features of miRNAs such as sequence, structure, and conservation in order to identify novel miRNA precursors as shown in Fig. 3.
2. SSCprofiler is trained to learn the characteristic features of human miRNA precursors with high accuracy, and the trained model is applied on CAGRs in search of novel miRNA genes.
3. Predictions are ranked according to expression information from a recently published full genome tiling array study (23), and the top four scoring candidates have been verified experimentally using Northern blot.

2.5.1.3. MiRRim – Terai et al. (20)

MiRRim is similar to SSCprofiler as it also uses an HMM algorithm that considers structure and conservation features for predicting novel miRNA genes. However, sequence information is not taken into consideration by the algorithm. The main conceptual differences between the two tools are provided in Table 1.

2.5.2. Bayes Classifiers

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a
### Table 1
Comparison of SSCprofiler and miRRim tools

<table>
<thead>
<tr>
<th>Biological features of miRNAs</th>
<th>SSCprofiler</th>
<th>miRRim</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negative data</strong></td>
<td>Selected from 3’-UTRs, and filtered according to conservation and a minimum energy score to ensure that they resemble true miRNAs in both structure and conservation</td>
<td>Randomly selected 200 nt-long genomic regions with different degrees of conservation. No requirements for resemblance with true miRNAs</td>
</tr>
<tr>
<td><strong>Sensitivity/specificity (validation set)</strong></td>
<td>HMM score: 3 Sens: 88.95% Spec: 84.16%</td>
<td>Sens: ~70% Spec: ~90%</td>
</tr>
<tr>
<td><strong>Generalization (blind test set)</strong></td>
<td>Identification of 219 previously unseen miRNAs with an accuracy of 72.15%</td>
<td>No evaluation of performance on a blind test set</td>
</tr>
<tr>
<td><strong>Scanning procedure</strong></td>
<td>104 nt sliding window, shifted 1 nt at a time for positive as well as negative data</td>
<td>Size of sliding window and shift step unclear. Positive data ranged between 160 and 236 nt, negative data were 200 nt, implying a larger window and thus a smaller search space</td>
</tr>
<tr>
<td><strong>Total number of hits</strong></td>
<td>For a coverage of 96.0% (HMM threshold 11), ~5,800 miRNA hits for 350 MB (CAGRs) of the human genome</td>
<td>For a coverage of 91.0%, ~4,000 miRNA hits for the whole human genome</td>
</tr>
<tr>
<td><strong>Expression information using high-throughput methods</strong></td>
<td>Tiling array data from HeLa and HepG2 cells</td>
<td>No expression information is provided</td>
</tr>
<tr>
<td><strong>Experimental verification</strong></td>
<td>Successful verification of four top scoring miRNA candidates via Northern blot</td>
<td>No experimental verification is provided</td>
</tr>
</tbody>
</table>

supervised learning setting. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

2.5.2.1. NaiveBayes – Yousef et al. (19)

1. This method generates the model automatically and identifies rules based on the miRNA gene sequence and structure; thus allowing the prediction of nonconserved miRNAs.
2. In addition, the method uses a comparative analysis over multiple species to reduce the false positive (FP) rate.
The miRDeep algorithm uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA (Solexa sequencing as well as 454 deep sequencing) with the secondary structure of the miRNA precursor.

The authors demonstrate the accuracy and robustness of the tool using published *C. elegans* data and data generated by deep sequencing human and dog RNAs.

miRDeep reports altogether ~230 previously unannotated miRNAs, of which 4 novel *C. elegans* miRNAs have been validated by Northern blot analysis.

SVMs are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an \( n \)-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets.

Multiple studies have utilized SVMs to predict novel miRNA genes, some of which are listed below:

1. Xue et al. (17) proposed a set of novel features of local contiguous structure–sequence information for distinguishing the hairpins of real pre-miRNAs and 1,000 pseudo pre-miRNAs. Remarkably, the SVM classifier built on human data can also correctly identify up to 90% of the pre-miRNAs from other species, including plants and virus, without utilizing any comparative genomics information.

2. Sewer et al. (34) focused on genomic regions around already known miRNAs in order to incorporate the observation that miRNAs are occasionally found in clusters. Starting with the known human, mouse, and rat miRNAs, the authors scanned 20 kb of flanking genomic regions for the presence of putative precursor miRNAs. Each genome was analyzed separately, allowing the evaluation of the species-specific identity and genome organization of miRNA loci. Only cross-species comparisons were used to make conservative estimates of the number of novel miRNAs. This ab initio method predicted between 50 and 100 novel pre-miRNAs for each of the considered species. Around 30% of these miRNAs have already been experimentally verified in a large set of cloned mammalian small RNAs (24).

In addition to miRNA gene prediction tools, a number of methods have been developed to complement such tools and help to maximize their prediction accuracy.

RNAmicro is designed specifically to work as a “sub-screen” for large-scale ncRNA surveys with RNAz. RNAz is an SVM-based
classification tool which combines comparative sequence analysis and structure prediction (35). RNAmicro tries to provide an annotation of the RNAz survey data, in order to provide a more balanced trade-off between sensitivity and specificity. Thus, the goal of RNAmicro is a bit different from that of specific surveys for miRNAs in genomic sequences; in the latter case, one is interested in very high specificity so that the candidates selected for experimental verification contain as few FPs as possible.

RNAmicro is an SVM-based method which, in order to classify the miRNA precursors, evaluates the information contained in multiple sequence alignment.

RNAmicro consists of the following components:

1. A pre-processor that performs the following three actions:
   a. Identifies conserved hairpin-structure regions in a multiple sequence alignment. Extracted windows of length $L$ are used in one-nucleotide steps from the input alignment.
   b. RNAalifold algorithm (Vienna RNA Package) is used to compute consensus sequence and structure (36).
   c. The consensus secondary structure which is obtained in “dot-parenthesis” notation is further analyzed.

2. A module that computes a vector of numerical descriptors from each hairpin structure.

3. A SVM that classifies the candidate based on its vector of descriptors.

Another study describes a SVM classifier that can separate between true and false Drosha processing sites. The biological relevance of this tool is based on the assumption that mature miRNAs are processed from long hairpin transcripts by Drosha, and that this processing defines the mature product and is characteristic for all miRNA genes. This classifier can predict the exact location of 5¢ microprocessor processing sites in human 5¢-miRNAs with 50% accuracy. It is also important to mention that if the predicted site is wrong then the actual site is within two nucleotides of the predicted site, in about 90% of the cases. Even though Microprocessor SVM is not effective as a stand-alone tool, it can be useful as:

1. A post-processor for existing tools that only predict whether hairpins are likely miRNAs.

2. A complementary miRNA gene classifier that performs better than currently available methods for predicting unconserved miRNAs. As a consequence, other prediction tools can be improved upon post-processing using this method.
In order to stress the importance of miRNA gene prediction tools in identifying miRNA genes implicated with various types of cancers, it is necessary to provide representative examples.

1. These include miRNAs predicted by sequence homology to already cloned miRNAs, such as miR-143 (25) involved in colorectal cancer (4), miR-125b (lin-4) and miR-145 which are implicated in breast cancer (37), miR-106a which is believed to play a regulatory role in colon, pancreas, and prostate cancer (38), and miR-155 which is associated with HL, BCL, pediatric BL, breast, and lung cancer as well as poor survival (8, 39–42).

2. Another large study utilized conservation with mouse and Fugu rubripes sequences and the score given by the program MiRscan (28) in order to predict novel miRNAs associated with cancer. These included mir221/222 which are involved in Papillary thyroid carcinoma (43) and glioblastomas (44), hsa-mir-192 which is shown to have reduced expression in colorectal neoplasia (4), hsa-mir-196a-1 which was cloned from human osteoblast sarcoma cells (45), and hsa-mir-210 which is implicated in Kaposi’s sarcoma-associated herpes virus infections (46). The majority of these miRNA genes were identified computationally and their implication in cancer confirmed by experimental methods.

3. In a recent large-scale bioinformatics study, Calin et al. (9) made use of the miRNA registry as well as bibliography for CAGRs, to show that over 80 known miRNAs reside within CAGR and fragile sites. This was one of the first large-scale computational studies to indicate a direct connection between genomic location of miRNAs and regions prone to genetic alteration in cancer. This study has provided the raw material for undergoing experiments aiming to verify these connections.

4. Recently, a sophisticated computational tool (SSCprofiler (21)) scanned over 70 CAGRs and resulted in the prediction of four novel candidate miRNAs, all of which were verified by tilling array data as well as Northern blot analysis.

This section underlies the capabilities and limitations of the numerous available computational methods for miRNA gene prediction.
Computational Identification of miRNAs Involved in Cancer

1. The general rule is that, the more biological information integrated in the computational tool the more successful it will be.

2. As indicated in Table 2, early tools making use of sequence and the closely related species conservation alone did not achieve very high prediction accuracy.

3. Subsequent methodologies taking into account additional biological information, such as structure and multiple species conservation for filtering their predictions, showed significant improvements.

4. Another important criterion for success is the simultaneous consideration of biological information. Considering all features at once, preferably at the nucleotide level, is more informative and more efficient than undertaking a pipeline approach, which utilizes the different features sequentially to predict novel miRNAs.

5. Sophisticated machine learning algorithms process all the biological features in parallel in order to build predictive models. This is a significant improvement to linear approaches adopted by initial brute-force methods (see Table 2).

6. Successful training of learning algorithms requires caution when selecting positive and negative training examples. Online databases may contain false-positives, and the definition of a negative miRNA is still uncertain.

7. The use of 3' UTR regions to draw negative miRNA genes has been the norm in most studies mostly because there was no documented miRNA gene within these regions. However, recently a small percentage of miRNA genes have been shown to exist within 3' UTRs.

8. Evaluating the performance of prediction tools by measuring sensitivity as well as specificity. However, these measures are greatly affected by the quality as well as the number of negative and positive samples (19). Hence, the prediction accuracy of one tool may change if the dataset from another study is used and vice versa. In general, computational miRNA gene prediction lacks a benchmark dataset.

9. There is still ample space for improvement in the field of computational prediction of miRNA genes, besides the great advances in the last years.

10. As more biological information regarding miRNA biogenesis and regulation is made available, computational tools incorporating this information will become much more effective.

11. Perhaps novel biological information will shed some light into one of the bottlenecks of in silico prediction of miRNAs, namely the identification of the mature miRNA sequence on the miRNA precursor.
<table>
<thead>
<tr>
<th>Features</th>
<th>Cloning</th>
<th>MiRscan</th>
<th>MiRseeker</th>
<th>Phylogenetic shadowing</th>
<th>Blat-ting</th>
<th>MiR Align</th>
<th>ProMir</th>
<th>Bayes-classifier</th>
<th>Xue</th>
<th>Sewer</th>
<th>RNA micro</th>
<th>SSC profiler</th>
<th>MiR Deep</th>
<th>MiR Rim</th>
<th>Microprocessor</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directly⁴</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirectly</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base pairing</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hairpin</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulges/loops</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature location</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermo-dynamic temp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conservation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conserved synteny</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Conserved clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple species</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brute-force</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homology based</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probablistics</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement other tools</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>74</td>
<td>75</td>
<td>73</td>
<td>97</td>
<td>93.30</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>96</td>
<td>91</td>
<td>88.10</td>
<td>64</td>
<td>84.16</td>
</tr>
<tr>
<td>Species</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nematode + human</td>
<td>Drosophila</td>
<td>Human</td>
<td>Mouse</td>
<td>Multi-species</td>
<td>Human, mouse, and rat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human</td>
<td>Mouse</td>
<td>Multi-species</td>
<td>Human, mouse, and rat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caenorhabditis elegans, human, and dog</td>
<td>Human</td>
<td>Human</td>
<td>Human</td>
</tr>
</tbody>
</table>

Use of large-scale, high-throughput data: x x

Experimental verification: x x x

Reference

*Directly — in the sense that the nucleotide distribution in the sequence is taken into consideration (i.e., GC content). Indirectly refers to the use of sequence to derive structure*
12. It is also important to mention that miRNA gene prediction is currently addressed as a 2D problem. Secondary structure prediction algorithms only follow 2D rules and do not portray a complete tertiary picture. Tools capable of predicting tertiary structure of miRNAs (such as pseudoknots (47)) will transform a 2D problem into a 3D one, reflecting the conditions found in the cell more accurately.

13. As a final note, it is important to stress that the development of computational tools is tightly linked to biological research. The successful evolution of these tools demands that developers keep up with novel biological findings that may change the way information should be used. A characteristic example of this is the use of Drosha processing sites in the Microprocessor (15) study mentioned above. It was recently shown that intronic microRNA precursors may bypass Drosha processing (48).

Acknowledgments

This work was supported by the action 8.3.1 (Reinforcement Program of Human Research Manpower – “PENED 2003” (03ED842)) of the operational program “competitiveness” of the Greek General Secretariat for Research and Technology, a Marie Curie Fellowship of the European Commission (PIOF-GA-2008-219622), and the National Science Foundation (NSF 0515357).

References


MicroRNA and Cancer
Methods and Protocols
Wu, W. (Ed.)
2011, X, 273 p., Hardcover
ISBN: 978-1-60761-862-1
A product of Humana Press