MicroRNAs and Cancer—The Search Begins!

Anastasis Oulas, Martin Reczko, and Panayiota Poirazi

***Abstract*—For almost three decades, cancer was thought to result from changes in the structure and/or expression of protein coding genes. The discovery of thousands of genes that produce noncoding RNA (ncRNA) transcripts in the past few years suggested that the molecular biology of cancer is much more complex. MicroRNAs (miRNAs), an important group of ncRNAs, have recently been associated with tumorigenesis by acting either as tumor suppressors or oncogenes. Experimental prediction of miRNA genes is a slow process, because of the difficulties of cloning ncRNAs. Complementary to experimental approaches, a number of computational tools trained to recognize features of the biogenesis of miRNAs have significantly aided in the prediction of new miRNA candidates. By narrowing down the search space, computational approaches provide valuable clues as to which are the dominant features that characterize these regulatory units and which genes are their most likely targets. Moreover, through the use of high-throughput expression profiling methods, many molecular signatures of miRNA deregulation in human tumors have emerged. In this review, we present an overview of existing computational methods for identifying miRNA genes and assessing their expression levels, and analyze the contribution of such tools toward illuminating the role of miRNAs in cancer.**

***Index Terms*—Cancer, computational prediction, microRNA genes.**

I. Introduction

MicroRNAs (miRNAs) belong to a recently identified group of the large family of noncoding RNAs [1]. The mature miRNA, usually 21–25 nt in length, is originally derived from a larger precursor, ~60–70 nt long, that folds into an imperfect stem-loop structure. In animals, these miRNA precursors (pre-miRNA) are themselves derived from cleavage of the primary miRNA (pri-miRNA) transcript by a multiprotein complex made up of Drosha RNase III [2] and Pasha (partner of Drosha), a double-stranded(ds) RNA-binding protein [3], [4]. After cleavage, miRNA precursors are transported into the cytoplasm by a cargo transporter called exportin 5 [5], [6]. Subsequently, the miRNA precursors are cleaved into an imperfect dsRNA duplex by another endonuclease RNase III enzyme, called Dicer [2], [7]. This duplex is composed of the mature miRNA strand and its complementary strand, commonly noted as miRNA*.

The mode of action of the mature miRNA in mammalian systems is dependent on complementary base pairing to the 3′-untranslated region (UTR) of the target mRNA, thereafter causing the inhibition of translation and/or the degradation of the mRNA. 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 let-7 [8], miR-15a/miR-16-1 cluster [9], and neighboring miR-143/miR-145 [10] are found to be reduced in some malignancies, suggesting their potential role as tumor suppressors. In contrast, some other miRNAs, such as the miR-17-92 cluster [11]–[13] and miR-155/BIC [14], 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) [15]. MiRNA genes located within, or in close proximity, to these regions have been suggested to be associated with chromosomal events leading to carcinogenesis (see Fig. 1). According to recent findings, miR-15a and miR-16a are located at chromosome 13q14, a region deleted in more than half of B cell chronic lymphocytic leukemias (B-CLLs) [17]. Detailed deletion and expression analysis showed that miR-15a and miR-16a are located within a 30-kb region of loss in B-CLL and that both genes are deleted or downregulated in the majority (~68%) of B-CLL cases [9]. Chromosome 13q14 deletions also occur in ~50% of mantle cell lymphoma, 16%–40% of multiple myeloma, and 60% of prostate cancers, suggesting that one or more tumor suppressor genes at 13q14 are involved in the pathogenesis of these human tumors [17].

About 30% of miRNAs are found in the introns[1] of other genes. It is generally thought that in most cases where the miRNA lies in the same orientation, the miRNA is cotranscribed with the host gene [18]. This can have obvious implications when investigating genes that are highly expressed in certain cancers, whereby the careful investigation of their intronic sequences can lead to the discovery of putative, intron-residing miRNAs with a fundamental role in cancer development.

In this review, we provide an overview of existing computational tools for the identification of novel miRNA genes and discuss the contribution of miRNA gene prediction toward understanding the biological background of cancer (schematically shown in Fig. 1). In Section II, we explain the need for fast and

Manuscript received April 10, 2008. Current version published January 4, 2009. This work was supported in part by the Action 8.3.1 (Reinforcement Program of Human Research Manpower—"PENED 2003") of the operational program ''competitiveness'' of the Greek General Secretariat for Research and Technology and in part by the INFOBIOMED European Network of Excellence (NoE) FP6-IST-2002-507585 European Union (EU) funded project.

A. Oulas is with the Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion GR-71110, Greece, and also with the Department of Biology, University of Crete, Heraklion 71409, Greece.

M. Reczko is with the Institute of Oncology, Biomedical Sciences Research Center (BSRC) Alexander Fleming, Athens 16672, Greece (e-mail: mreczko@googlemail.com).

P. Poirazi is with the Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion GR-71110, Greece (e-mail: poirazi@imbb.forth.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2008.2007086

[1]Noncoding sections of DNA

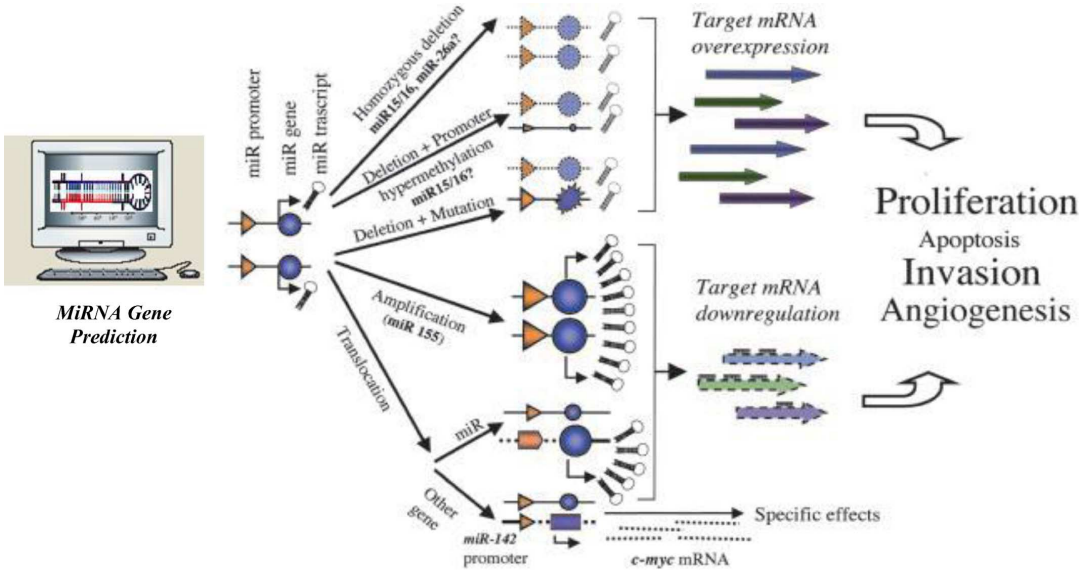

Fig. 1. MiRNAs as cancer players. Computational prediction initiates the search for putative miRNAs that play a role in tumorigenesis. Some of these proposed mechanisms are experimentally proven, like the deletion of *miR-15a/miR-16a* cluster in B-CLL [9], [16], the c-myc overexpression by the reposition near a putative *miR* promoter [15], or *miR143/miR-145* cluster downregulation in colon cancers [10]. Figure adopted from Callin *et al.* [15].

cost-effective identification of novel miRNAs. In Section III, we present the characteristics of miRNAs that have to be considered in order to construct computational models that will successfully predict new miRNAs. Section IV provides a historical overview of the computational tools available to date while, Section V provides a comparison of these tools and a summary of the criteria for success. Finally, we suggest ways for further improving the computational prediction of miRNA genes in the future and discuss their contribution to the identification of novel cancer players.

II. Identifying Novel miRNAs

Recent estimates show that over 30% of vertebrate genomes is transcribed (expressed) [19], and out of this, only 1% is coding genes, while the rest has to be various types of noncoding RNA genes. Based on these findings, a vast amount of unexplored noncoding areas are likely to exist in the human genome. Presently, only ~700 human miRNA hairpin sequences are listed in the miRNA registry (version 12), most of which have been experimentally verified, and it is anticipated that there may be thousands more. In addition, the recent discovery of a number of miRNA genes involved in various cancers [9], [14], [20]–[22] has steered the interest of a large portion of the scientific community and the identification of novel miRNAs and associated targets has become a very important and rapidly growing field. In order to identify novel miRNA genes from different organisms, flexible, reliable, and fast prediction methods are required. Experimental prediction of miRNAs and their target interaction is currently cumbersome. Some miRNAs are difficult to isolate by cloning due to low expression, stability, tissue specificity, and technical difficulties of the cloning procedure while selecting the right genomic region to investigate is often a very challenging task of its own. Furthermore, as shown from computational as well as experimental evidence, miRNAs usually target more than one gene, while target genes are frequently regulated by multiple miRNAs [23]–[25]. This combinatorial mode of regulation makes experimental characterization of miRNA and target interactions very difficult and time-consuming. Computational prediction of miRNA genes from genomic sequences offers a much faster, cheaper, and effective way of identifying putative miRNA genes and potential targets. Moreover, by predicting the location of miRNA genes, these methods enable experimentalists to concentrate their efforts on genomic regions more likely to contain novel miRNA genes, thus facilitating the discovery process.

The tools and algorithms for the computational prediction of ncRNAs have been thoroughly reviewed previously [26]. In this paper, we provide a review of the tools and algorithms that have been specifically employed for computational prediction of miRNAs. We further discuss the potential contributions of these predictions toward deciphering the molecular biology of cancer and the design of new prognostic tools.

III. Computational Prediction Using miRNA Sequence, Structure, and Conservation

Accurate prediction of new miRNAs requires the consideration of certain characteristic properties of miRNAs based on either experimental [27]–[29] or computational evidence [30]–[34] that can be used to build a classification scheme or predictive model. These general features include sequence composition, secondary structure, and species conservation.

Initial approaches for computational prediction of miRNA genes focused on simple sequence similarity searches, using pairwise sequence alignment algorithms such as basic local alignment tool (for nucleotides) (BLASTN) [35], in an attempt to find possible homologues and orthologues of already cloned miRNAs. This technique led to the discovery of several new miRNAs [36], [37]; however, it failed to detect miRNAs that lacked a certain degree of sequence similarity due to species

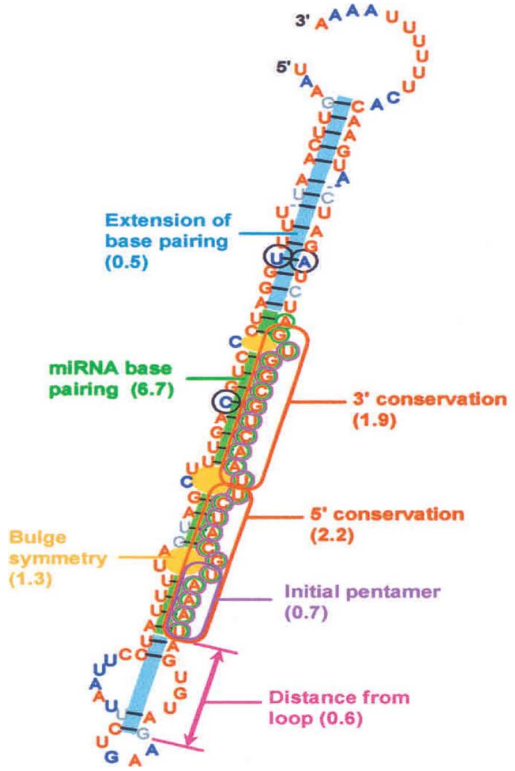

Fig. 2. Criteria used by MiRscan to identify miRNA genes among aligned segments of two genomes. The typical hairpin-like structure of miRNAs and the seven components of the MiRscan score for mir-232 of *C. elegans*/*C. briggsae*. Figure adopted from Lim *et al*. [32].

divergence. In order to overcome this limitation, prediction methods made use of RNA secondary structure folding algorithms. MiRNAs have a very characteristic hairpin-like (stem-loop) secondary structure (see Fig. 2) in their precursor form. Programs such as MFOLD [38] and RNAfold [39] predict RNA secondary structure by free-energy minimization and have been utilized in many studies in order to assess whether a certain genomic sequence (in an RNA form) has the capacity to fold into the hairpin-like structure typical of miRNAs. Another characteristic feature of most miRNAs is their profound conservation across multiple different species (i.e., mouse and human), even highly divergent ones (chicken and fugu). In general, regions of the genome that are functional are expected to be evolutionary conserved across species, while regions that do not exhibit any profound biological function should be less conserved. This feature has significantly contributed to the prediction of novel miRNAs using information of multiple sequence alignments of whole genomes from a variety of different organisms. The general observation is an island of conservation in a region of otherwise unconserved genomic sequence denoting the presence of an miRNA precursor (a phenomenon later referred to as phylogenetic shadowing).

IV. Computational Tools for miRNA Prediction

A. Sequence-Based Prediction

Early prediction methods relied on sequence conservation with already cloned miRNAs to predict novel miRNA genes. For example, Quintana *et al.* [36] described the prediction of 34 novel miRNAs by tissue-specific cloning of approximately 21-nt RNAs from mice. Almost all identified miRNAs were found to be conserved in the human genome and are also frequently found in non-mammalian vertebrate genomes, such as pufferfish. In heart, liver, or brain, it was found that a single, tissue-specifically expressed miRNA dominates the population of expressed miRNAs, suggesting a role for these miRNAs in tissue specification or cell lineage decisions.

B. Sequence, Structure, and Closely Related Species Conservation

The initial computational tools for miRNA gene prediction used simple rules derived from sequence and structural features of miRNA precursors. Conservation has also been used as a feature but was limited to pairwise comparisons of closely related organisms. Two representative approaches include the MiRscan [32] and the miRseeker [40] tools.

MiRscan implements a supervised method whereby known miRNAs are used as a training set to predict for novel miRNA candidates. The method was applied to search for hairpin structures using RNAfold (free energy <-25 kcal/mol) within sequences that are conserved between *C. elegans* and *C. briggsae* (Washington University (WU)-basic local alignment search tool (BLAST) cutoff $E < 1.8$). At first, 36 000 hairpins were predicted including 50/53 miRNAs previously reported to be conserved between the two species. These 50 miRNAs were used as a training set for the development of an MiRscan program. MiRscan was then used to evaluate the 36 000 hairpins.

The MiRscan algorithm searches for several features of the hairpin in a 21-nt sliding window that is shifted across the hairpin. These features include: base-pairing, 3′-conservation, 5′-conservation, bulge symmetry, distance from the loop, and initial pentamer properties, as illustrated in Fig. 2. The total score for an miRNA candidate is computed by summing the weighted score of each feature.

A three-part computational pipeline called miRseeker was used for predicting miRNA genes in two *Drosophila* species: *D. melanogaster* and *D. pseudoobscura*. The two Drosophila genomes were first aligned to establish conservation. MiRseeker, which includes MFOLD, was then utilized to identify Drosophila miRNA sequences. The algorithm's efficiency was assessed by observing its ability to give high scores to 24 known Drosophila miRNAs. A key feature of miRseeker is the consideration of pattern divergence as the authors detected less selective pressure in the loop sequences of orthologous precursor miRNAs.

The initial results concerning the conservation of miRNA genes motivated researchers to further investigate the role of this feature in miRNA gene prediction. This led to the development of methods that utilized multiple, rather than pairwise, sequence comparisons of closely related species. Furthermore, later approaches began to investigate the conservation in the genomic location of miRNAs across different organisms. Two of these approaches are described shortly.

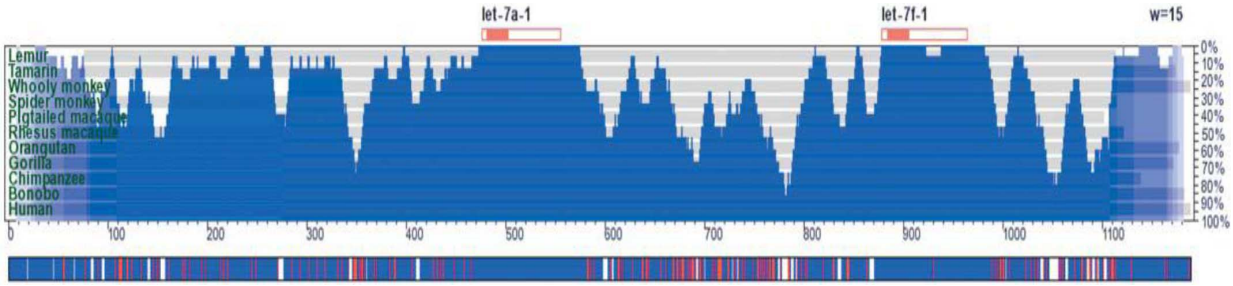

Fig. 3. Prediction of novel miRNA genes using phylogenetic shadowing profiles representation of phylogenetic shadowing results in ten primate species for a genomic region harboring two known miRNA genes. Pre-miRNAs and mature miRNAs are represented as light gray (or red) open and solid boxes on the top of the figure, respectively. Intensity of blue indicates sequence coverage depth and background horizontal gray rectangles show the coverage of individual monkey species. The bar below the alignment represents the nature of observed variations: dark gray (or blue)—no variation, light gray (or red)—substitution, white—insertion/deletion. Figure adopted from Berezikov *et al*. [41].

C. Multiple Species Conservation and Conserved Regions of Synteny

A strategy known as phylogenetic shadowing [41] is another approach used to predict novel human miRNAs. This approach is based on observing an island of conservation in a region of otherwise unconserved genomic sequence denoting the presence of an miRNA precursor. It considers multiple, rather than pairwise, sequence comparison of closely related species and provides an accurate method for identifying conserved regions down to the nucleotide level. Comparison of sequences in [41] from more than 100 miRNA regions in ten different primates revealed a characteristic profile: variation in loop sequences, conservation in stems of hairpins, and a significant decrease in conservation of sequences flanking the hairpins. This pattern was used to identify potential new miRNAs in pairwise alignments of more divergent species, such as human and mouse or human and rat. After additional filtering, such as looking at the folding free energy of candidate sequences, 976 potential human miRNAs were identified. This set contained over 80% of all known human miRNAs in release version 3.1 of the miRNA Registry (http://microrna.sanger.ac.uk/sequences/). Northern blot analyses combined with database searches reached a conservative estimate of 200–300 verified novel human miRNAs, a twofold increase over previous studies [29]. Fig. 3 shows an example of how phylogenetic shadowing profiles over ten primate species can be used to identify new miRNAs. Strong conservation over all species is evident only for two well-known miRNA genes.

In another approach, the sequence comparison tool BLAST-like alignment tool (BLAT) [42] was 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/). This was one of the first approaches to make use of full genome sequence alignments, the incentive being that human and mouse miRNAs should reside in conserved regions of synteny[2] [37]. Results were further filtered using secondary structure prediction tools (MFOLD) and other criteria (G:U base pairings). This method made use of the characteristic feature of some miRNAs to cluster together and show conservation in the location of clustering with respect to the genes around them. The findings of this paper included the prediction of 80 new putative miRNA genes (35 human and 45 mouse 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.

D. Towards a More General Model for miRNA Prediction

One of the first tools that implemented nonconservation features is miRAlign [43]. This tool detects new miRNAs based on both sequence and structure alignment. Two main characteristics make miRAlign distinct from existing homologue search methods: first, the ability to find distant homologs does not depend on sequence conservation of the whole pre-miRNA sequence nor the nearly perfect match of the $\sim$22 nt mature part, but a relatively loose conservation of the mature miRNA sequence. Second, more properties of miRNA structure conservation are considered. And unlike profile search methods, which need relatively large family members to construct the profile, miRAlign introduces a structure alignment strategy and can use each single miRNA as a query to do homology search. In comparison with other tools, miRAlign was found to outperform conventional BLAST search and easy RNA profile identification (ERPIN) search [44] by achieving higher sensitivity and comparable specificity. A main advantage of this method is the prediction of more distant miRNA homologs or orthologs.

E. Next Generation of Tools

More sophisticated prediction tools make use of machine learning algorithms to detect miRNA genes and employ a more general approach whereby all three features of sequence, structure, and conservation are combined in a way that allows the prediction of miRNAs on a full genome scale and is not limited to closely related species. These methods use true miRNAs and "negative" samples to train and construct a model with

[2]In other words, miRNAs of humans and those of other closely related organisms should be located in areas surrounded by homologous genes of similar function

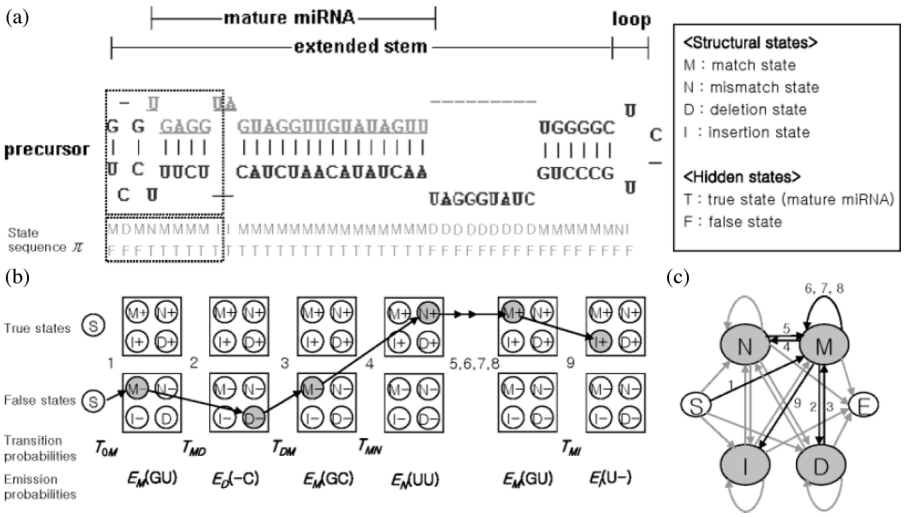

Fig. 4. Pairwise representation of stem–loop structures and state sequences of pre-miRNAs, where the state of each pair includes structural information and mature miRNA region information (hidden states). (a) Structure of the pre-miRNA with the mature miRNA region shown in light gray. (b) Transition and emission scheme of the structural states and the hidden states for pairwise sequence in the dotted rectangle shown in (a). (c) Four-state finite-state automaton. Figure adopted from Nam *et al*. [45].

optimized sensitivity and specificity. Negative samples are usually hairpin-like sequences derived from 3′-UTRs, an area of the genome with no documented miRNA genes. Next, we discuss two such methods that utilize a probabilistic approach.

Nam *et al.* [45] constructed a highly specific probabilistic colearning method based on the paired hidden Markov model (HMM) to identify close homologs as well as distant homologs. The topology and states of the HMM are handcrafted based on prior knowledge and assumptions, and the probabilities are derived from the data. This method combines both sequence and structural characteristics of miRNA genes in a probabilistic framework, and simultaneously decides if an miRNA gene and the mature miRNA are present by detecting the signals for the site cleaved by Drosha. As shown in Fig. 4, each paired nucleotide position on the pre-miRNA is represented using hidden states (true or false) with positions on the mature miRNA denoted as true. Structural information describing whether the paired nucleotides are in a match or mismatch state is also included. Nam *et al.* [45] employed this HMM to predict novel miRNA genes and finally filtered their candidates using conservation across multiple divergent species. The authors validated their candidates by examining the accumulation of pri-miRNAs in the cells depleted of Drosha.

In another interesting approach [34], Yousef *et al*. used a simple naive Bayes classifier to predict miRNA genes. Given an example $X = (x_1, \ldots, x_n)$, where x_i denote the different miRNA features, the method searched for a class C that maximizes the likelihood: $P(X|C) = P(x_1, \ldots, x_n|C)$. The (naive Bayes) assumption of conditional independence among the features, given the class, allows expression of this conditional probability $P(X|C)$ as a product of simpler probabilities

$$P(\mathrm{X}|C) = \prod_{i=1}^{n} P(x_i|C). \tag{1}$$

This method generates the model automatically and identifies rules based on the miRNA gene sequence and structure, thus allowing prediction of nonconserved miRNAs. In addition, the method uses a comparative analysis over multiple species to reduce the false positive (FP) rate. This allows for a tradeoff between sensitivity and specificity. Based on findings over multiple genomes, this method appears to be applicable to a wide variety of eukaryotes. The resulting algorithm demonstrates higher specificity and similar sensitivity compared to currently available algorithms, which use conserved genomic regions to reduce false positives [32], [40], [46]. Like Nam *et al*. [45], rather than relying on miRNAs homology between related species, Yousef *et al*. [34] directly use features of the miRNA sequence and secondary structure. However, in contrast to Nam *et al.* [45], the naive Bayes classifier is trained to identify miRNAs directly from the data, rather than handcrafting a model. In this study, prior knowledge is used for initial filtering of the data, but not for constructing the model. The naive Bayes classifier is a standard model with no domain-specific assumptions (apart from the usual conditional independence assumptions inherent to the model). In addition, whereas Nam's model was trained and tested on a single type of data (136 human miRNAs) with respect to a restricted set of negative examples, in the Yousef study, the model was trained and tested using a variety of miRNAs from multiple organisms. Yousef *et al*. demonstrate that the results obtained by training and testing, using arbitrarily selected numbers of negative samples, are highly sensitive to the size of the negative set. Hence, to overcome this problem, they use multiple sources of miRNAs. By integrating data from multiple species, it appears that the learning process is stabilized. The explanation for this is that distribution of the positive data used for training the classifier better represents the variety of miRNA classes. It is clear that by providing more examples for training and testing, the classifier demonstrates better generalization. Moreover, a model is constructed that is more likely to be applicable to a variety of genomes.

Another class of supervised algorithms that have been commonly used for miRNA gene prediction are support vector machines (SVMs) [30], [31], [33], [47], [48]. Xue *et al*. [48]

proposed a set of novel features of local contiguous structure–sequence information for distinguishing the hairpins of real pre-miRNAs and 1000 pseudo pre-miRNAs. SVMs were applied on these features to classify real versus pseudo pre-miRNAs, achieving approximately 90% accuracy on human data. 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.

Various different approaches base their methodology on the different features and characteristics of miRNAs. An interesting case is the study by Sewer *et al*. [33], which focuses 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 (pre-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 [49].

F. Methods Designed to Complement Other Tools

In addition to the numerous stand-alone tools for miRNA gene prediction, a number of programs have been developed to complement and refine the results of existing software.

The program RNAmicro [31] is designed specifically to work as a "subscreen" for large-scale ncRNA surveys with RNAz [50]. The goal of RNAmicro is thus 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 false positives as possible. RNAmicro, in contrast, tries to provide an annotation of the RNAz survey data, in order to provide a more balanced tradeoff between sensitivity and specificity similar to that of annotating protein motifs in predicted protein coding genes.

RNAmicro presents an SVM-based classification for microRNA precursors that is designed to evaluate the information contained in multiple sequence alignments. It consists of: 1) a preprocessor that identifies conserved "almost-hairpins" in a multiple sequence alignment; this is done by extracting windows of length L in 1-nt steps from the input alignment. For each window, consensus sequence and consensus structure are computed using the RNAalifold algorithm [39] implemented in the Vienna RNA package [39]. The automaton is then used to analyze the consensus secondary structure, which is obtained in "dot-parenthesis" notation; 2) A module that computes a vector of numerical descriptors from each "almost-hairpin;" and 3) a support vector machine used to classify the candidate based on its vector of descriptors.

Based on the fact that mature miRNAs are processed from long hairpin transcripts by Drosha, another study describes an SVM classifier that can separate between true and false processing sites [30]. Even though it is only the first of several steps, the initial Drosha processing defines the mature product and is characteristic for all miRNA genes. Methods that can separate between true and false processing sites are therefore essential to miRNA gene discovery.

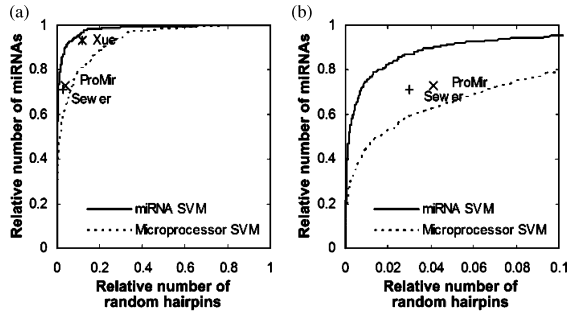

Fig. 5. Microprocessor processing site prediction improves miRNA gene prediction of other SVM tools. (a) Dotted line shows the receiver operating characteristic (ROC) curve for the microprocessor SVM predictor that was trained on true versus false processing sites in pri-miRNAs. The solid line shows the ROC curve for the microprocessor SVM predictor that was trained on miRNAs versus random hairpins on top of the results from other prediction tools. True positives are estimated based on cross-validation; false positives are based on the genomic predictions. The points for Sewer, ProMir, and Xue are the sensitivities and specificities reported by Sewer *et al*. [33], Nam *et al*. [45], and Xue *et al*. [48]. (b) Detailed excerpt of (a). Figure adopted from Xue *et al*. [48].

This later study [30] describes an SVM classifier that can predict 5′ microprocessor processing sites in human 5′ miRNAs with 50% accuracy. A microprocessor site corresponds to the position where cleavage is presumed to take place by Drosha. Importantly, if the predicted site is wrong, the actual site is within 2 nt of the predicted site in about 90% of the cases. This microprocessor SVM can be useful as a postprocessor for existing tools that only predict whether hairpins are likely miRNAs. Furthermore, microprocessor processing site predictions can be used to create another miRNA gene classifier that performs better than currently available methods for predicting unconserved miRNAs. As shown in Fig. 5, microprocessor is not very effective as a stand-alone tool; however, other prediction tools are improved upon postprocessing using microprocessor.

Since microprocessor SVM predicts the 5′ processing site, predictions of the 3′ processing site—and as a consequence, 3′ miRNAs—will be less accurate. Nevertheless, the second miRNA gene classifier performance is independent of whether the mature miRNA is from the 5′ or 3′ stem. By using the two classifiers to analyze 130 recently published miRNA sequences [51], this study further shows that several of these sequences do not share the characteristics of previously known miRNAs. Importantly, the lack of common characteristics as measured by these classifiers correlates with a lack of evidence for the reported sequences being miRNAs. This correlation suggests that current databases may contain falsely annotated miRNAs.

V. Tool Comparison and Criteria for Success

Given the availability of numerous methods for miRNA gene prediction, a comparison of their capabilities and limitations is much needed. In general, the success degree of such tools is

TABLE I
COMPARISON OF 12 miRNA PREDICTION TOOLS

Features	Cloning	MiRscan	miRseeker	phylogenetic shadowing	Blatting	miRAlign	ProMir	Bayes-classifier	Xue	Sewer	RNAmicro	Microprocessor SVM
Sequence												
Directly[a]	X					X	X	X	X	X	X	X
Indirectly		X	X	X								
Structure												
Base pairing		X				X	X	X	X	X	X	X
Hairpin		X				X	X	X	X	X	X	X
Bulges/Loops		X				X	X	X	X	X	X	X
Mature location		X					X					
Thermodymamic temp			X	X	X	X	X	X	X	X	X	X
Conservation												
Pairwise	X	X	X									
Conserved synteny					X							
Conserved clustering										X		
Multiple species				X			X	X				
Methodology												
Bruteforce		X	X									
Homology based				X	X							
SVM									X	X	X	X
Probablistics						X	X	X				
Complement other tools											X	X
Performance												
sensitivity		0.74%	75%				73%	97%	93.3%	64%	90%	90%
specificity							96%	91%	88.1%	64%		78%
species		Human + Nematode	Drosophila				human	mouse	Multi species	Human,mouse, rat	human	human
Relevant Reference		[32]	[40]	[41]	[42]	[43]	[45]	[34]	[48]	[33]	[31]	[30]

Table shows the included in each tool, the performance achieved, and the species specificity.
[3] 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.

largely dependent on the biological information that goes into building them. As shown in Table I, initial tools utilizing solely sequence and closely related species conservation did not have very high prediction accuracy. Subsequent methodologies that also took into consideration biological information such as structure and multiple species conservation for filtering their predictions resulted in significant improvements. The simultaneous consideration of this information is another important criterion for success. Considering all of these features at once is more informative than undertaking a pipeline approach, whereby the different features are used sequentially to predict novel miRNAs. A number of sophisticated machine learning algorithms are now used to build models trained to recognize all of these features in parallel rather than the linear approach adopted by initial brute-force methods (see Table I). For the successful training of these algorithms, special care must be taken when selecting positive and negative training examples since online databases may contain false positives and the definition of a negative miRNA remains unclear. Most papers use 3′-UTR regions to draw their negatives. However, the fact that no miRNA has been documented to exist within these regions does not mean that such an observation will not occur in the future. Furthermore, sensitivity and specificity are directly affected by the number as well as the quality of negative and positive samples [34], so initial results from one study may change if the dataset from another study is used and vice versa. Table I summarizes the optimal performance achieved by 12 miRNA prediction tools as well as the species specificity of each tool.

While great advances have been made over the last decade in computational identification of miRNA genes, there is plenty of space for improvement. Computational tools will become much more accurate when more biological information regarding miRNA biogenesis and regulation is available. One of the bottlenecks in *in silico* prediction of miRNAs is the identification of the mature miRNA sequence on the miRNA precursor. The small size of the mature (22 nt) limits the sequence, structure, and conservation information available within and around this region. Improvement will also result from the incorporation of novel features such as the flanking region around the miRNA gene and what information it can provide as to hint for the presence of an miRNA gene in the vicinity. We believe that a more complete picture will emerge as tools that are capable of predicting tertiary structure of miRNAs (such as pseudoknots, [52]) become available.

This will transform a 2-D problem into a 3-D one, reflecting more accurately the conditions found in the cell. As a final note to this section, it is important to mention that the development of these tools is tightly linked to biological research. The successful evolution of these tools demands that developers keep track of novel biological findings that often change the way information should be used. A characteristic example of this is the use of Drosha processing sites in the microprocessor [30] study mentioned earlier. It was recently shown that intronic microRNA precursors may bypass Drosha processing [53].

VI. DISCUSSION

A. Role of MicroRNAs in Cancer

The unique biogenesis of miRNAs as well as their characteristic features (sequence structure and conservation) differentiates them from other ncRNAs and facilitates their computational prediction. Due to their implication in several diseases, including cancer, these tiny molecules are already the target of intensive research aiming at novel pharmacological interventions. Computational tools are thus very useful for the efficient and fast prediction of novel miRNAs as well as their targets.

Multiple miRNAs associated with various cancers have been identified by computational methods. Representative examples that were predicted by sequence homology to already cloned mouse miRNAs include the *miR-143* [36] involved in colorectal

cancer [10], *miR-125b (lin-4)* and *miR-145* that are implicated in breast cancer [54], *miR-106a* that is believed to play a regulatory role in colon, pancreas and prostate cancer [55], and *mir-155* that is associated with Hodgkin lymphoma (HL), B cell lymphomas (BCL), pediatric Burkitt lymphoma (BL), breast, and lung cancer as well as poor survival [14], [56]–[59]. Furthermore, a large study predicted multiple vertebrate miRNA genes using conservation with mouse and Fugu rubripes sequences [29] and the score given by MiRscan. MiRNAs predicted in this study include *mir221/222* that are involved in papillary thyroid carcinoma [60] and glioblastomas [61], *hsa-mir-192* that is shown to have reduced expression in colorectal neoplasia [10], *hsa-mir-196a-1* that was cloned from human osteoblast sarcoma cells [62], and *hsa-mir-210* that is implicated in Kaposi's sarcoma-associated herpes virus infections [63].

According to a large-scale bioinformatics study by Calin *et al*. that made use of information from online databases, more than 50 known human miRNAs reside within CAGRs and FRA sites [15]. These findings indicate a connection between the genomic location of miRNAs and regions prone to alteration in cancer. Many of these connections as well as other associations between miRNAs and cancer, including multiple computationally predicted miRNAs, have been verified experimentally. Some of this experimental evidence is reviewed shortly.

As mentioned previously, one of the first experimentally verified connections between CAGRs and miRNAs concerned the *mir-15a* and *mir-16-1* cluster that is located within a minimal region of loss of heterozygosity (LOH) at the 13q14.3 locus, a region commonly deleted in B-CLLs [9]. Similar results from another study provided experimental evidence that 28 miRNAs are differentially expressed in colonic adenocarcinoma when compared with normal mucosa [10]. Among these, *mir-143* and *mir-145* consistently displayed reduced steady-state levels in colorectal cancer.

Further support for the role of miRNAs in cancer came from a study utilizing array-based comparative genomic hybridization (aCGH) to screen different types of cancer for the presence of miRNA alterations. For the 283 known miRNAs subjected to this analysis, 37.1% were found to exhibit DNA copy number alterations in ovarian cancer, 72.8% in breast cancer, and 85.9% in melanoma, displaying an overall distinct miRNA genomic alteration pattern for each tumor type [22]. This research provided further evidence that miRNAs could potentially act as oncogenes or tumor suppressors.

The let-7 family members are another example of miRNAs that reside within a fragile site [15] implicated in lung cancer [8]. Experimental evidence shows that *let-7* miRNAs are significantly downregulated in lung cancer, and when they are ectopically overexpressed in lung adenoma cell lines, they are able to inhibit cell growth, indicating that they can function as tumor suppressors [8]. Furthermore, the expression levels of the *let-7* family members in lung cancer were shown to be negatively correlated with RAS, a very well known oncogene, indicating a possible repression of RAS by these miRNAs [64].

A well-known genomic alteration in human cancer is the frequently amplified 13q31 locus in lymphoma. The oncogene involved in this type of cancer was unidentified until scientists looked closer at the noncoding C13orf25 gene that was shown to contain the large *mir-17-92* miRNA cluster. He *et al.* [60] later found that the expression of six miRNAs was significantly correlated with the gene dosage of C13orf25 [12]. Five of these miRNAs belonged to the *mir-17-92* cluster. Indeed, the precursor of *mir-27-92* is upregulated more than fivefold in $\sim$65% of B cell lymphoma. In parallel with this study, an independent group also found that the *mir-17-92* cluster is involved in B cell lymphoma. The miRNA expression profile analysis of a human B cell line revealed that the miRNAs of the *mir-17-92* cluster are significantly upregulated when *c-myc* (a very powerful oncogene) is turned on [65].

The overexpression of *mir-155*, which also resides in a fragile site, was first reported in children with BL [14]. Subsequent studies showed that it is also upregulated in BCL and HL [56], [58], as well as breast [54] and colon carcinoma [55]. *Mir-155* is transcribed together with the noncoding gene *bic*, which has been shown to be overexpressed and promote B cell lymphoma in cooperation with *myc* [66], [67].

Despite the body of evidence supporting a role of miRNAs in cancer, their exact mechanism of action remains to be elucidated. Toward this goal, miRNA target prediction tools can offer a first indication as to which target genes are regulated by miRNAs, thus providing new insights regarding their specific functions and guiding future experiments.

Experimental verification of computational predictions will be the ultimate step in revealing the molecular pathways of miRNA regulation and characterizing their involvement in cancer.

B. Expression Profiling for miRNAs

Expression profiling analysis of protein-coding genes has already provided deep insights into cancer biology and cancer diagnosis. High-throughput profiling methods are also increasingly used for genome-wide assessment of expression profiles of mature miRNAs. Using microarrays, a recent study showed that miRNAs often exhibit tissue-specific expression signatures, and that these signatures may be implicated in tissue differentiation [68]. As shown in Fig. 6(a), expression profiling analysis revealed the differential expression of miRNAs across different types of cancers. Interestingly, these alterations are more frequently associated with downregulation than upregulation of miRNAs. Since miRNAs are believed to regulate tissue differentiation [68], this association may reflect differentiation defects commonly found in cancer cells. The hierarchical clustering in Fig. 6(a) also shows that specific signatures may exist for each cancer type, suggesting that miRNA may be used for cancer classification and perhaps prognosis prediction. Lu *et al*. [21] compared expression profiles of miRNAs and mRNAs, and showed that miRNA expression profiles are superior for tumor-type classification of poorly differentiated cancers [see Fig. 6(b)]. This observation further validates the inherent relationship of miRNAs with differentiation state as well as developmental lineage.

A more general role for miRNAs and cancer was recently suggested by Vilonia *et al*. In this study, the expression profile of 228 miRNAs in 540 samples of six different solid tumors

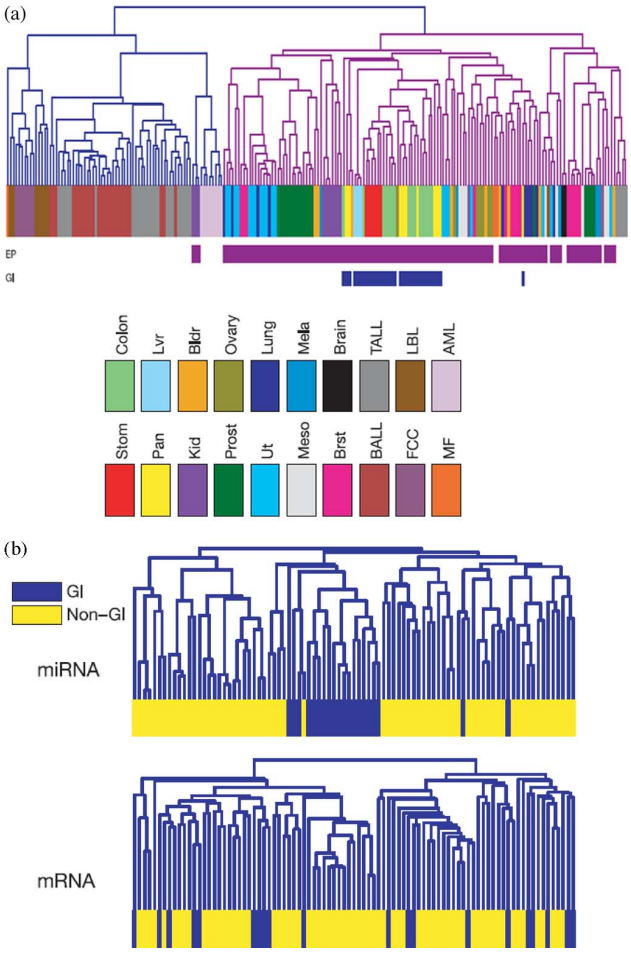

Fig. 6. (a) Hierarchical clustering of miRNA expression. miRNA profiles of 218 samples from several different cancer tissues (denoted by the gray-scale or color scheme) shows that tumors of the same cancer can be grouped together based on miRNA gene expression profiling (average linkage, correlation similarity). Samples are in columns, miRNAs in rows. Samples of epithelial (EP) origin or derived from the gastrointestinal tract (GI) are indicated. (b) Comparison of miRNA data and mRNA data. For 89 epithelial samples from (a) that had mRNA expression data, hierarchical clustering was performed. Samples of GI origin are shown in dark gray (or blue). GI-derived samples largely cluster together in miRNA expression space, but not in mRNA expression space. Figure adopted from Lu *et al*. [21].

(lung, gastric, colon, breast, prostate, and pancreas) was analyzed. Results revealed a common signature composed of 21 miRNAs that are differentially expressed in at least three tumor types [55]. The list of differentially expressed miRNAs contained well-characterized cancer-associated miRNAs, including *miR-17-5p*, *miR-20a*, *miR-21*, *miR-92*, *miR-106a*, and *miR-155*. One miRNA, *mir-21*, was overexpressed in all six types and two, *mir-17–5p* and *mir-191*, were overexpressed in five types of cancer. Overexpression of *mir-21* in glioblastoma may promote cell growth, and thus, a more malignant phenotype by blocking expression of critical apoptosis-related genes [69].

In contrast to the report by Lu *et al*. [21], most of the miRNAs identified in Vilonia study showed upregulation in cancers. It is currently unclear whether this apparent discrepancy resulted from the use of different technical platforms or differences in the sample numbers and tissues analyzed. The fact that some of the miRNA levels were high in all or most of these tumors implies a general role in oncogenesis.

C. Concluding Remarks

The amount of work supporting a pivotal role of miRNAs in cancer is rapidly growing over the last few years, indicating that these molecules may soon become the focus of cancer-related research. Computational miRNA gene prediction tools can provide invaluable information regarding the location of putative miRNA genes in the genome, thus guiding and complementing the slow experimental procedures commonly used. As a cautionary note, we should stress that for any computational model to prove its value, the predicted miRNA genes or at least a sample of highest scoring candidates should be experimentally verified. Unfortunately, this combined approach has only been used in a couple of studies [41], [45], mostly because experimental verification involves technical problems that can be difficult to solve. An example is whether or not the predicted miRNA gene candidate is expressed in the tissue culture used to perform the experiments, an issue that is currently being addressed by *tilling arrays* [70]. As more and more miRNAs are discovered, high-throughput methods such as microarrays will allow for a global, genome-wide expression profile of miRNAs. Recent microarray experiments, using various tumor tissues, hint to either a down-regulation of miRNAs, alluding to a tumor suppressor role, or an upregulation of miRNAs, indicating an oncogenic function. Future experiments are expected to further characterize the tissue specificity of particular miRNAs, provide unique miRNA gene signatures that define certain cancer types, and open up new avenues for cancer prognosis and treatment.

REFERENCES

[1] R. Lee, R. Feinbaum, and V. Ambros, "The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[2] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V. N. Kim, "The nuclear RNase III Drosha initiates microRNA processing," *Nature*, vol. 425, no. 6956, pp. 415–419, 2003.

[3] A. M. Denli, B. B. Tops, R. H. Plasterk, R. F. Ketting, and G. J. Hannon, "Processing of primary microRNAs by the microprocessor complex," *Nature*, vol. 432, no. 7014, pp. 231–235, 2004.

[4] R. I. Gregory, K. P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar, "The microprocessor complex mediates the genesis of microRNAs," *Nature*, vol. 432, no. 7014, pp. 235–240, 2004.

[5] E. Lund, S. Guttinger, A. Calado, J. E. Dahlberg, and U. Kutay, "Nuclear export of microRNA precursors," *Science*, vol. 303, no. 5654, pp. 95–98, 2004.

[6] Y. Qin, R. Yi, I. G. Macara, and B. R. Cullen, "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs," *Genes Dev.*, vol. 17, no. 24, pp. 3011–3016, 2003.

[7] G. Hutvágner and P. D. Zamore, "A microRNA in a multiple-turnover RNAi enzyme complex," *Science*, vol. 297, no. 5589, pp. 2056–2060, 2002.

[8] J. Takamizawa, H. Konishi, K. Yanagisawa, S. Tomida, H. Osada, H. Endoh, T. Harano, Y. Yatabe, M. Nagino, Y. Nimura, T. Mitsudomi, and T. Takahashi, "Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival," *Cancer Res.*, vol. 64, no. 11, pp. 3753–3756, 2004.

[9] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C. M. Croce, "Frequent deletions and down-regulation of

micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 24, pp. 15524–15529, 2002.

[10] M. Z. Michael, S. M. O'Connor, N. G. van Holst Pellekaan, G. P. Young, and R. J. James, "Reduced accumulation of specific microRNAs in colorectal neoplasia," *Mol. Cancer Res.*, vol. 1, no. 12, pp. 882–891, 2003.

[11] Y. Hayashita, H. Osada, Y. Tatematsu, H. Yamada, K. Yanagisawa, S. Tomida, Y. Yatabe, K. Kawahara, Y. Sekido, and T. Takahashi, "A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation," *Cancer Res.*, vol. 65, no. 21, pp. 9628–9632, 2005.

[12] L. He, J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond, "A microRNA polycistron as a potential human oncogene," *Nature*, vol. 435, no. 7043, pp. 828–833, 2005.

[13] H. Tagawa and M. Seto, "A microRNA cluster as a target of genomic amplification in malignant lymphoma," *Leukemia*, vol. 19, no. 11, pp. 2013–2016, Nov. 2005.

[14] M. Metzler, M. Wilda, K. Busch, S. Viehmann, and A. Borkhardt, "High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma," *Genes Chromosomes Cancer*, vol. 39, pp. 167–169, 2003.

[15] G. A. Calin, C. Sevignani, C. D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, and C. M. Croce, "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 9, pp. 2999–3004, 2004.

[16] A. Cimmino, G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini, and C. M. Croce, "miR-15 and miR-16 induce apoptosis by targeting BCL2," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 13944–13949, 2005.

[17] F. Bullrich and C. M. Croce, *Chronic Lymphoid Leukemia*. New York: Marcel Dekker, 2001.

[18] S. Baskerville and D. P. Bartel, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes," *RNA*, vol. 11, no. 3, pp. 241–247, Mar. 2005.

[19] C. Fantom, "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, Sep. 2005.

[20] G. A. Calin, C.-G. Liu, C. Sevignani, M. Ferracin‡, N. Felli, C. D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono, M. L. Dell'Aquila, H. Alder, L. Rassenti, T. J. Kipps, F. Bullrich, M. Negrini, and C. M. Croce, "MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 32, pp. 11755–11760, 2004.

[21] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub, "MicroRNA expression profiles classify human cancers," *Nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[22] L. Zhang, J. Huang, N. Yang, J. Greshock, M. S. Megraw, A. Giannakakis, S. Liang, T. L. Naylor, A. Barchetti, M. R. Ward, G. Yao, A. Medina, A. O'Brien-Jenkins, D. Katsaros, A. Hatzigeorgiou, P. A. Gimotty, B. L. Weber, and G. Coukos, "MicroRNAs exhibit high frequency genomic alterations in human cancer," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 24, pp. 9136–9141, 2006.

[23] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou, "A combined computational-experimental approach predicts human microRNA targets," *Genes Dev.*, vol. 18, no. 10, pp. 1165–1178, 2004.

[24] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[25] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nat. Methods*, vol. 3, no. 11, pp. 881–886, Nov. 2006.

[26] B.-J. Yoon and P. P. Vaidyanathan, "Computational Identification and Analysis of noncoding RNAs," *IEEE Signal Process. Mag.*, vol. 24, no. 1, pp. 64–74, Jan. 2007.

[27] A. Khvorova, A. Reynolds, and S. D. Jayasena, "Functional siRNAs and miRNAs exhibit strand bias," *Cell*, vol. 115, no. 2, pp. 209–216, Oct. 2003.

[28] K. Jeon, Y. Lee, J. T. Lee, S. Kim, and V. N. Kim, "MicroRNA maturation: Stepwise processing and subcellular localization," *EMBO J.*, vol. 21, no. 17, pp. 4663–4670, 2002.

[29] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel, "Vertebrate microRNA genes," *Science*, vol. 299, no. 5612, pp. 1540–1540, 2003.

[30] S. A. Helvik, O. Snove, Jr., and P. Saetrom, "Reliable prediction of Drosha processing sites improves microRNA gene prediction," *Bioinformatics*, vol. 23, no. 2, pp. 142–149, 2006.

[31] J. Hertel and P. F. Stadler, "Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data," *Bioinformatics*, vol. 22, pp. e197–e202, 2006.

[32] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, and S. Yekta, "The microRNAs of *caenorhabditis elegans*," *Genes Dev.*, vol. 17, no. 8, pp. 991–1008, 2003.

[33] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M. J. Brownstein, T. Tuschl, E. van Nimwegen, and M. Zavolan, "Identification of clustered microRNAs using an ab initio prediction method," *BMC Bioinf.*, vol. 6, pp. 267–281, 2005.

[34] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, "Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier," *Bioinformatics*, vol. 22, pp. 1325–1334, 2006.

[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[36] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, "Identification of tissue-specific microRNAs from mouse," *Current Biol.*, vol. 12, pp. 735–739, 2002.

[37] M. J. Weber, "New human and mouse microRNA genes found by homology search," *FEBS J.*, vol. 272, no. 1, pp. 59–73, 2005.

[38] M. Zuker and D. H. Mathews, "Turner algorithms and thermodynamics for RNA secondary structure prediction: A practical guide," in *RNA Biochemistry and Biotechnology*, J. Barciszewski and B. F. C. Clark, Eds. NATO ASI Series, Norwell, MA: Kluwer Academic, 1999.

[39] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucl. Acids Res.*, vol. 31, pp. 3429–3431, 2003.

[40] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin, "Computational identification of Drosophila microRNA genes," *Genome Biol.*, vol. 4, no. 7, pp. R42–R61, 2003.

[41] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, R. H. Plasterk, and E. Cuppen, "Phylogenetic shadowing and computational identification of human microRNA genes," *Cell*, vol. 120, pp. 21–24, 2005.

[42] W. J. Kent, "BLAT–the BLAST-like alignment tool," *Genome Res.*, vol. 12, pp. 656–664, 2002.

[43] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang, and Y. Li, "MicroRNA identification based on sequence and structure alignment," *Bioinformatics*, vol. 21, pp. 3610–3614, 2005.

[44] A. Lambert, J. F. Fontaine, M. Legendre, F. Leclerc, E. Permal, F. Major, H. Putzer, O. Delfour, B. Michot, and D. Gautheret, "The ERPIN server: An interface to profile-based RNA motif identification," *Nucl. Acids Res.*, vol. 32, pp. W160–W165, 2004.

[45] J. W. Nam, K. R. Shin, J. Han, Y. Lee, V. N. Kim, and B. T. Zhang, "Human microRNA prediction through a probabilistic co-learning model of sequence and structure," *Nucl. Acid Res.*, vol. 33, no. 11, pp. 3570–3581, 2005.

[46] Y. Grad, J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim, "Computational and experimental identification of *C. elegans* microRNAs," *Mol. Cell*, vol. 11, no. 5, pp. 1253–1263, 2003.

[47] A. H. Buck, J. Santoyo-Lopez, K. A. Robertson, D. S. Kumar, M. Reczko, and P. Ghazal, "Discrete clusters of virus-encoded micrornas are associated with complementary strands of the genome and the 7.2-kilobase stable intron in murine cytomegalovirus," *J. Virol.*, vol. 81, pp. 13761–13770, 2007.

[48] C. Xue, F. Li, T. He, G. P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinf.*, vol. 6, pp. 310–316, 2005.

[49] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R. U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H. I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl, "A mammalian microRNA expression atlas based on small RNA library sequencing," *Cell*, vol. 129, pp. 1401–1414, 2007.

[50] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding RNAs," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 2454–2459, 2005.
[51] J. M. Cummins, Y. He, R. J. Leary, R. Pagliarini, L. A. Diaz, Jr., T. Sjoblom, O. Barad, Z. Bentwich, A. E. Szafranska, E. Labourier, C. K. Raymond, B. S. Roberts, H. Juhl, K. W. Kinzler, B. Vogelstein, and V. E. Velculescu, "The colorectal microRNAome," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 3687–3692, 2006.
[52] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinf.*, vol. 5, pp. 104–115, 2004.
[53] J. G. Ruby, C. H. Jan, and D. P. Bartel, "Intronic microRNA precursors that bypass Drosha processing," *Nature*, vol. 448, pp. 83–86, 2007.
[54] M. V. Iorio, M. Ferracin, C. G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Menard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini, and C. M. Croce, "MicroRNA gene expression deregulation in human breast cancer," *Cancer Res.*, vol. 65, pp. 7065–7070, 2005.
[55] S. Volinia, G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce, "A microRNA expression signature of human solid tumors defines cancer gene targets," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 2257–2261, 2006.
[56] P. S. Eis, W. Tam, L. Sun, A. Chadburn, Z. Li, M. F. Gomez, E. Lund, and J. E. Dahlberg, "Accumulation of miR-155 and BIC RNA in human B cell lymphomas," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 3627–3632, 2005.
[57] J. Kluiver, E. Haralambieva, D. de Jong, T. Blokzijl, S. Jacobs, B. J. Kroesen, S. Poppema, and A. Van Den Berg, "Lack of BIC and microRNA miR-155 expression in primary cases of Burkitt lymphoma," *Genes Chromosomes Cancer*, vol. 45, pp. 147–153, 2006.
[58] J. Kluiver, S. Poppema, D. de Jong, T. Blokzijl, G. Harms, S. Jacobs, B. J. Kroesen, and A. Van Den Berg, "BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas," *J. Pathol.*, vol. 207, pp. 243–249, 2005.
[59] N. Yanaihara, N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, G. A. Calin, C. G. Liu, C. M. Croce, and C. C. Harris, "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis," *Cancer Cell*, vol. 9, pp. 189–198, 2006.
[60] H. He, K. Jazdzewski, W. Li, S. Liyanarachchi, R. Nagy, S. Volinia, G. A. Calin, C. G. Liu, K. Franssila, S. Suster, R. T. Kloos, C. M. Croce, and A. de la Chapelle, "The role of microRNA genes in papillary thyroid carcinoma," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 19075–19080, 2005.
[61] S. A. Ciafre, S. Galardi, A. Mangiola, M. Ferracin, C. G. Liu, G. Sabatino, M. Negrini, G. Maira, C. M. Croce, and M. G. Farace, "Extensive modulation of a set of microRNAs in primary glioblastoma," *Biochem. Biophys. Res. Commun.*, vol. 334, pp. 1351–1358, 2005.
[62] M. Lagos-Quintana, R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl, "New microRNAs from mouse and human," *RNA*, vol. 9, pp. 175–179, 2003.
[63] X. Cai, S. Lu, Z. Zhang, C. M. Gonzalez, B. Damania, and B. R. Cullen, "Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 15, pp. 5570–5575, 2005.
[64] S. M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K. L. Reinert, D. Brown, and F. J. Slack, "RAS is regulated by the let-7 microRNA family," *Cell*, vol. 120, no. 5, pp. 635–647, 2005.
[65] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell, "c-Myc-regulated microRNAs modulate E2F1 expression," *Nature*, vol. 435, no. 7043, pp. 839–843, 2005.
[66] B. E. Clurman and W. S. Hayward, "Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: Evidence for stage-specific events," *Mol. Cell. Biol.*, vol. 9, no. 6, pp. 2657–2664, 1989.
[67] W. Tam, D. Ben-Yehuda, and W. S. Hayward, "Bic, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA," *Mol. Cell. Biol.*, vol. 17, no. 3, pp. 1490–1502, 1997.
[68] C. G. Liu, G. A. Calin, B. Meloon, N. Gamliel, C. Sevignani, M. Ferracin, C. D. Dumitru, M. Shimizu, S. Zupo, M. Dono, H. Alder, F. Bullrich, M. Negrini, and C. M. Croce, "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 26, pp. 9740–9744, 2004.
[69] J. A. Chan, A. M. Krichevsky, and K. S. Kosik, "MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells," *Cancer Res.*, vol. 65, no. 14, pp. 6029–6033, 2005.
[70] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermuller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras, "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, 2007.

Anastasis Oulas received the B.Sc. degree (with honors) in molecular genetics in biotechnology from the Department of Biological Sciences, University of Sussex, Falmer, U.K., in 2001, and the M.Sc. degree (with distinction) in computational genetics and bioinformatics from the Imperial College, London, U.K., in 2002. He is currently working toward the Ph.D. degree at the Computational Biology Laboratory, Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion, Greece. He is also enrolled in the Postgraduate Program Molecular Biology and Biomedicine at the University of Crete, Heraklion.

In January 2003, he was a Research Assistant at the Computational Biology Laboratory IMBB-FORTH.

Martin Reczko received the Ph.D. degree in computer science from the University of Stuttgart, Stuttgart, Germany, in 1995.

He conducted a collaborative project with the German Cancer Research Center, Heidelberg, Germany, concerning the development of artificial neural networks for modeling systems in molecular biology. He has been the Chief Scientific Officer of Synaptic Ltd., a company he cofounded focusing on the development and application of machine learning and evolutionary methods in computational molecular biology. He was a Postdoctoral Researcher at the Democritus University of Thrace, Xanthi, Greece, where he developed advanced signal processing for magnetic resonance imaging. For several years, he has been the Principal Researcher—equivalent to a Research Associate Professor—of the bioinformatics activity at the Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece. Recently he joined the Biomedical Sciences Research Center "Alexander Fleming," Athens, Greece, as a Visiting Bioinformatics Researcher. His current research interests include the development of novel computational methods for modeling complex biological systems. He has authored or coauthored and reviewed for many journals and contributed in several national and European R&D projects.

Panayiota Poirazi was born in Cyprus, in 1974. She received the Diploma in mathematics (with honors) from the University of Cyprus, Nicosia, Cyprus, in 1996, and the M.S. and Ph.D. degrees in biomedical engineering from the University of Southern California (USC), Los Angeles, in 1998 and 2000, respectively.

She is currently a Research Associate Professor in computational biology in the Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology-Hellas (FORTH), Heraklion, Greece. Her current research interests include the area of computational modeling of biological systems with emphasis in the fields of neuroscience and functional genomics.

Prof. Poirazi was awarded a Marie Curie Fellowship from the European Commission (EC) for conducting research in the field of bioinformatics in January 2002. In 2005, she was awarded the European Molecular Biology Organization (EMBO) Young Investigator Award and in 2008, she was awarded a Second Marie Curie Fellowship for obtaining experimental training in behavioral and cellular neuroscience at the University of California, Los Angeles (UCLA).