BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2226–2234
doi:10.1093/bioinformatics/btq329

Genome analysis Advance Access publication June 30, 2010

MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data

Anthony Mathelier[1,2] and Alessandra Carbone[1,2,*]

[1]UPMC Univ. Paris 06, FRE3214, Génomique Analytique, 15 rue de l'Ecole de Médecine and [2]CNRS, FRE3214, Laboratoire de Génomique des Microorganismes, F-75006 Paris, France

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are a class of endogenes derived from a precursor (pre-miRNA) and involved in post-transcriptional regulation. Experimental identification of novel miRNAs is difficult because they are often transcribed under specific conditions and cell types. Several computational methods were developed to detect new miRNAs starting from known ones or from deep sequencing data, and to validate their pre-miRNAs.

Results: We present a genome-wide search algorithm, called MIReNA, that looks for miRNA sequences by exploring a multidimensional space defined by only five (physical and combinatorial) parameters characterizing acceptable pre-miRNAs. MIReNA validates pre-miRNAs with high sensitivity and specificity, and detects new miRNAs by homology from known miRNAs or from deep sequencing data. A performance comparison between MIReNA and four available predictive systems has been done. MIReNA approach is strikingly simple but it turns out to be powerful at least as much as more sophisticated algorithmic methods. MIReNA obtains better results than three known algorithms that validate pre-miRNAs. It demonstrates that machine-learning is not a necessary algorithmic approach for pre-miRNAs computational validation. In particular, machine learning algorithms can only confirm pre-miRNAs that look alike known ones, this being a limitation while exploring species with no known pre-miRNAs. The possibility to adapt the search to specific species, possibly characterized by specific properties of their miRNAs and pre-miRNAs, is a major feature of MIReNA. A parameter adjustment calibrates specificity and sensitivity in MIReNA, a key feature for predictive systems, which is not present in machine learning approaches. Comparison of MIReNA with miRDeep using deep sequencing data to predict miRNAs highlights a highly specific predictive power of MIReNA.

Availability: At the address http://www.ihes.fr/~carbone/data8/

Contact: alessandra.carbone@lip6.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 8, 2010; revised on June 10, 2010; accepted on June 11, 2010

1 INTRODUCTION

MicroRNAs (miRNAs) are a class of endogenes of 18–25 nt in length and they occur in a precursor (pre-miRNA) with a characteristic hairpin secondary structure. miRNAs are involved in post-transcriptional regulation of protein-coding genes in animals and plants by sequence complementarity within the corresponding messenger RNA. Since the discoveries of *lin-4* and *let-7* (Lee *et al.*, 1993; Pasquinelli *et al.*, 2000) in *Caenorhabditis elegans*, the number of miRNA/pre-miRNA pairs found in miRBase (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006), the reference database collecting all known miRNAs from 115 species, has grown to more than 10000 (11488 in version 14.0). Experimental identification of novel miRNAs is difficult because they might be transcribed under specific conditions and cell types. Only the half (6166 out of 11488) of miRBase pre-miRNA/miRNA pairs have been experimentally validated.

Several computational methods were developed to detect miRNAs in genomes (see Tanzer *et al.*, 2008, for a review). A first family of methods tests specific physical and geometrical characteristics of the associated pre-miRNAs (Dezulian *et al.*, 2006; Hertel *et al.*, 2006; Legendre *et al.*, 2005; Weber, 2005); these properties concern the hairpin structure of sequences that are similar to experimentally known pre-miRNA sequences. A second family is constituted by *de novo* methods that use thermodynamic stability of pre-miRNA hairpins and characteristic patterns of sequence conservation (Hertel and Stadler, 2006; Lai *et al.*, 2003; Lim *et al.*, 2003a; Wang *et al.*, 2005) to predict pre-miRNAs. To confirm pre-miRNA sequences and their structures, tools that are based either on machine-learning algorithms (Jiang *et al.*, 2007; Sewer *et al.*, 2005; Xue *et al.*, 2005) or on ranking euclidean distances in a multidimensional space constructed from more than 30 parameters (Xu *et al.*, 2008) have been developed. A third approach to miRNA prediction is based on deep sequencing data. A computational tool, named miRDeep, predicting pre-miRNA/miRNA pairs by combining experimental data to a Dicer processing model has been introduced (Friedländer *et al.*, 2008).

We propose a new tool, named MIReNA, designed for answering several questions on miRNAs and pre-miRNAs. MIReNA searches for miRNAs and pre-miRNAs at genome scale, and computationally validates pre-miRNAs. It can be used in four different ways, depending on available data: known miRNAs, deep sequencing reads, potential miRNAs occurring in long sequences and putative pre-miRNAs containing potential miRNAs. It has been compared to four existing computational tools: MiPred (Jiang *et al.*, 2007) that classifies sequences from real or pseudo pre-miRNAs by using a random forest prediction model with combined features, miR-abela (Sewer *et al.*, 2005) that uses an *ab initio* approach to identify clustered miRNAs based on the validation of pre-miRNAs by a

*To whom correspondence should be addressed.

machine learning method, microPred (Batuwita and Palade, 2009), a classifier system for real and pseudo pre-miRNAs that uses machine learning techniques, miRDeep (Friedländer *et al.*, 2008) that predicts pre-miRNA/miRNA pairs from deep sequencing data.

2 THE ALGORITHM

MIReNA is a search algorithm, which is designed for the prediction of miRNAs and pre-miRNAs. It can be used for a genome-wide miRNA or pre-miRNA discovery. To identify pre-miRNA/miRNA pairs, it explores a multidimensional space defined by only five (physical and combinatorial) parameters. These parameters characterize suitable pre-miRNA structures and allow the identification of a strikingly simple set of conditions that make MIReNA powerful at least as much as more sophisticated computational methods, such as machine learning algorithms, used to confirm pre-miRNA sequences (Jiang *et al.*, 2007; Sewer *et al.*, 2005; Xue *et al.*, 2005).

2.1 Five numerical criteria to describe a pre-miRNA

We call miRNA* the complementary sequence r^* (possibly including unpaired nucleotides) of the miRNA r within a pre-miRNA structure. Given a pre-miRNA sequence s, the *adjusted minimum folding energy* (*AMFE*(s); Zhang *et al.*, 2006) is computed as $\frac{\text{MFE}(s)}{l(s)} * 100$, where MFE($s$) stands for the minimum free energy of s (Delisi and Crothers, 1971; Tinoco *et al.*, 1971) and $l(s)$ is the length, i.e. the number of nucleotides, of s. The *minimum free energy index* (*MFEI*(s); Zhang *et al.*, 2006) is computed as $\frac{\text{AMFE}(s)}{\%GC}$, where *%GC* stands for the percentage of G+C composition of s. A pre-miRNA secondary structure is defined to satisfy five main properties, the first three fixing the combinatorial structure of the pre-miRNA and the last two its physico-chemical characteristics:

(I) a miRNA r cannot fold with itself within the pre-miRNA secondary structure;

(II) the inequalities $0.83 \leq \frac{l(r^*)}{l(r)} \leq 1.17$ hold. To compute the thresholds, we considered the distance between r and r^* to be defined as $|l(r)-l(r^*)|/l(r)$, where $l(r), l(r^*)$ are the lengths of r, r^*, respectively. A distance close to zero implies a strong proximity of the sequences with a matching of essentially all nucleotides in r, r^*; larger distances imply a larger number of unpaired nucleotides between the sequences. By analyzing the distribution of distances between a miRNA and its miRNA* on the testing set, we estimated an acceptable distance to be $< \mu + \sigma$, where μ, σ are mean and SD of the distribution. This corresponds to define an interval of acceptable ratios to be $1-(\mu+\sigma)$ and $1+(\mu+\sigma)$, where a ratio $= 1$ corresponds to a matching between sequences of the same length. The testing set is defined by the 6166 experimentally verified miRNAs in miRBase (version 14.0, all 115 species confounded) from which pre-miRNA/miRNA pairs with the miRNA folding with itself or with the ratio $l(r^*)/l(r) > 3$ were removed;

(III) the percentage $p(r)$ of unmatched nucleotides in r within the pre-miRNA secondary structure is $\leq 26\%$ of $l(r)$. This threshold corresponds to $\mu+\sigma$, where μ, σ are mean and SD of the distribution of percentages computed on the testing set in II.

(IV) the pre-miRNA sequence has AMFE ≤ -32. This threshold corresponds to $\mu+\sigma$, where μ, σ are mean and SD of the distribution of AMFE values computed on the set of distinguished pre-miRNAs in miRBase containing an experimentally verified miRNA. pre-miRNA/miRNA pairs with the miRNA folding with itself or with the ratio $l(r^*)/l(r) > 3$ were removed from miRBase;

(V) the pre-miRNA sequence has MFEI < -0.85. This MFEI criterion was used in (Zhang *et al.*, 2006) to discriminate pre-miRNAs from other RNA sequences with high specificity. The threshold was found to sharply select only pre-miRNA sequences out of coding and non-coding RNA sequences.

Criteria II and III ask r, r^* to match almost completely: the lengths of r and r^* are comparable by II, and a high percentage of matched nucleotides is demanded by III. The five numerical criteria are used to identify suitable pre-miRNA structures. They allow us to consider structures with multiple stems, which are not accepted as feasible by many known approaches like MiPred for instance (Supplementary Fig. 1c and d).

2.2 MIReNA algorithm

MIReNA can handle four kinds of data, is characterized by specific pre-treatments of these data and predicts miRNAs and pre-miRNAs after using Criteria I–V to filter the associated structures. It handles known miRNAs, deep sequencing data, potential miRNAs occurring in long sequences and putative pre-miRNAs containing potential miRNAs. The first two kinds of data are checked against full genome sequences.

The first kind of large scale analysis takes into account a set of known miRNAs. Given a known miRNA, the algorithm searches for highly similar sequences in a genome. The outcoming sequences are potential miRNAs that are checked to have a suitable pre-miRNA structure surrounding them. The black path in Figure 1 illustrates the successive steps of the algorithm:

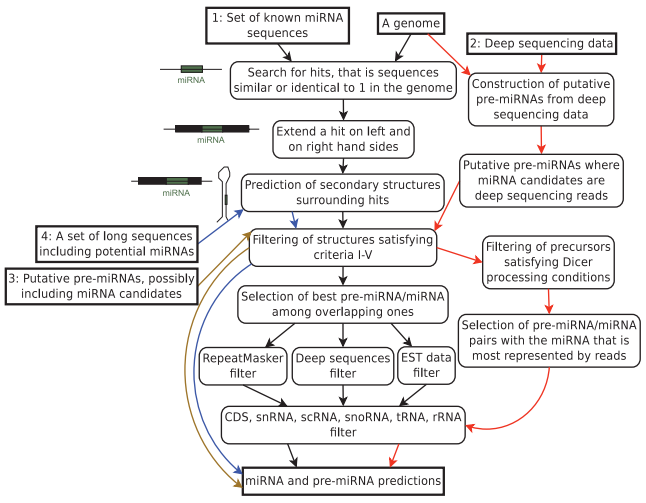

Fig. 1. Schema of MIReNA algorithm. MIReNA can be used in four different ways described as colored paths in the schema (black, red, brown and blue). Input and output data are detailed in squared boxes.

(1) it looks for sequences that are similar to the known miRNAs by minimizing the number of insertions, deletions and substitutions of nucleotides. Sequences obtained with this search are considered as potential miRNAs. An approximate string matching algorithm, i.e. an adaptation of (Myers, 1999), extracts, for each genome position, the minimal (in length) suffix sequence of at most 25 nt in length that displays at most 15% of errors (by the Levenshtein distance; Levenshtein, 1966) with the original miRNA sequence;

(2) the resulting sequences are extended on the left and on the right by 200 nt on each side;

(3) for each sequence extension obtained in Step 2, we compute secondary structures of all subsequences containing the potential miRNA. To produce secondary structures, we used an adapted implementation of RNAfold (Hofacher *et al.*, 1994);

(4) all structures obtained in Step 3 are scanned to filter out those that do not satisfy numerical Criteria I–V above. The remaining structures are considered as putative pre-miRNAs. (Notice that the sequences that we retain are at least 60-nt long.);

(5) we consider all potential miRNAs that are included in putative pre-miRNA sequences validated in Step 4. If two miRNAs overlap, we keep the corresponding pre-miRNA that has the best MFEI and discharge the most energetically unfavorable one. If two miRNAs r_1 and r_2 are contained in pre-miRNAs with the same MFEI and $p(r_1) < p(r_2)$, we keep r_1. If $p(r_1) = p(r_2)$ and $l(r_1^*)/l(r_1) < l(r_2^*)/l(r_2)$, we keep r_1. Otherwise, we keep the pre-miRNA whose miRNA has maximum length;

(6) (this step is optional) the resulting pre-miRNAs are filtered and either repeated sequences (RepeatMasker is used for this), or sequences that are not overlapping deep sequencing reads, or sequences that are not overlapping expressed sequence tags (ESTs) are discarded.

(7) (this step is optional) the resulting pre-miRNAs coming from either Step 5 or 6 are filtered using genomic information. Pre-miRNAs overlapping CoDing Sequences (CDS), scRNA, snRNA, snoRNA, tRNA, rRNA and 21U-RNA sequences on either strains are discarded.

The second kind of large scale analysis considers deep sequencing data (red path in Fig. 1). Following the same approach introduced in miRDeep (Friedländer *et al.*, 2008), we map deep sequencing reads on the genome and identify clustered regions of reads (at most 30 nt away) by discarding those which are >140 nt. Such sequences are considered as putative pre-miRNAs and all reads contained in them are taken as potential miRNAs. Secondary structures of putative pre-miRNAs are constructed using RNAfold. Step 4 above is run. It is followed by a step that filters precursors satisfying a Dicer processing condition and at least one condition among three concerning reads mapping the precursor. Among several miRNAs within the same precursor, we select the one which is most represented by reads. The screening based on genomic information might also be run.

MIReNA can realize Step 4 alone to check whether a putative pre-miRNA including a potential miRNA satisfies Criteria I–V (brown path in Fig. 1). Finally, it can predict pre-miRNAs included in sequences of about 400 nt in length by starting from a potential miRNA in each sequence. This is done by running Steps 3 and 4 (blue path in Fig. 1).

3 RESULTS

MIReNA has been tested in several manners. To validate Criteria I–V, we fully varied the corresponding thresholds, we evaluated different combinations of these criteria and we compared I–V to MiRfold criteria (Billoud *et al.*, 2005). MIReNA performance in the validation of pre-miRNAs (brown path in Fig. 1) has been evaluated against three predictive systems, miR-abela, MiPred and microPred. MIReNA predictive power on deep sequencing data (red path in Fig. 1) was compared to miRDeep. MIReNA was used to predict miRNAs and pre-miRNAs on six different species (black path in Fig. 1) starting from known miRNAs in miRBase.

3.1 Validation of Criteria I–V

MIReNA has been validated on several datasets of known pre-miRNA structures coming from the phylogenetically distant genomes of *Arabidopsis thaliana*, *C.elegans*, *Homo sapiens*, *Oryza sativa* and *Rattus norvegicus*. To demonstrate that the method confirms sequences as pre-miRNAs and does it with good specificity and sensitivity, we traced the receiver operating characteristic (ROC) curves on the datasets for the five different species varying systematically parameters. We studied ROC curves by varying all five criteria (see Fig. 2a and c, and Supplementary Fig. 3a, c, e, g, i, k, m and o), by fixing the combinatorial Criteria I–III and by varying the physico-chemical Criteria IV and V (see Fig. 2b and d, and Supplementary Fig. 3b, d, f, h, j, l, n and p), and by varying criteria I–III without considering Criteria IV and V (Supplementary Fig. 5). When using one criteria at the time, Criteria V is the one

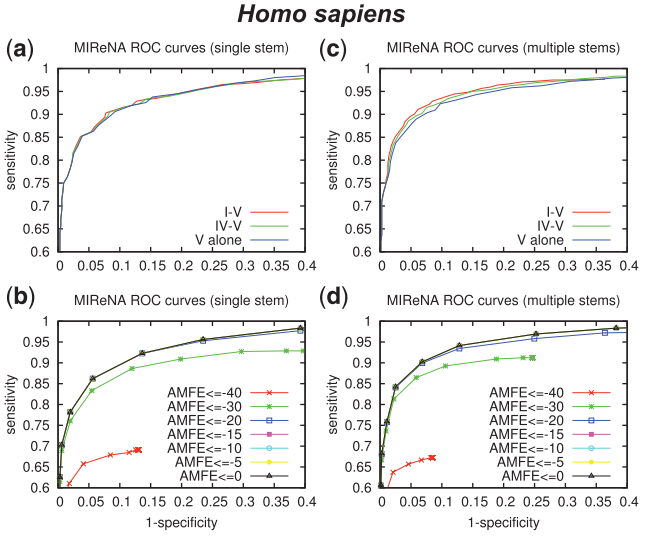

Fig. 2. Zoom on ROC curves for *H. sapiens*. Plots are constructed by applying MIReNA on datasets containing single stem (**a**) and (**b**), and multiple stem (**c**) and (**d**) pre-miRNA and pre-miRNA-like structures. (**a**) and (**c**) ROC curves have the best MCC obtained by varying different numerical criteria (see Supplementary Table 1). (**b**) and (**d**) ROC curves are obtained by varying Criteria IV–V. See Supplementary Figure 2 for complete ROC curves.

that provides best predictions and for this, we checked whether the coupling with the other criteria could improve its strength or not (see Fig. 2a and c, Supplementary Fig. 3a, c, e, g, i, k, m and o and Supplementary Table 1).

For all species, we used single and multiple stems datasets. MIReNA's ROC curves show a successful discrimination of real pre-miRNAs from pseudo ones on *H.sapiens* (Fig. 2a–d), *O.sativa*, *C.elegans*, *A.thaliana* and *R.norvegicus* (Supplementary Fig. 3). MIReNA's ROC curves approximate well with the maximal performance point [i.e. the top left corner in the plots, corresponding to an ideal curve passing through it and characterizing 100% of the area under the curve (AUC), or equivalently a 100% sensitivity and 100% specificity] as shown in Supplementary Table 1, where different numerical criteria provide a AUC >97% on all species. AUC, accuracy (Acc) and Mathew's correlation coefficient (MCC) values reported in Supplementary Table 1 correspond to the ROC curve with best MCC value. These curves are plot in Figure 2a and c and Supplementary Figure 3a, c, e, g, i, k, m and o.

By varying Criteria IV–V and letting fixed I–III, we observe that best ROC curves are comparable with those obtained by varying I–V (see Fig. 2 and Supplementary Fig. 3). Essentially, the same AUC values are obtained by different combinations of numerical criteria, but slightly improved Acc and MCC values are identified when the five criteria are considered all together, both for single and multiple stems validation. We explicitly measured the predictive power of Criteria V by ignoring the other criteria and letting V to vary. While best AUC values are reached by considering V alone, best Acc and MCC values are obtained with the combination of all criteria (see Supplementary Table 1). As expected, by varying combinatorial Criteria I–III and ignoring physico-chemical properties IV and V, ROC curves show a much lower discrimination power (see Supplementary Fig. 5).

MIReNA's ROC curves are stable for a large interval of AMFE values: curves computed for AMFE ≥ -20 overlap each other and for AMFE ≤ -30 the signal starts to degrade with a loss of sensitivity (and an improved specificity). This stability is an important feature for a user since it simplifies considerably parameterization. See Figure 2b and d, and Supplementary Figure 3b, d, f, h, j, l, n and p, where the AUC of MIReNA's ROC curves shows to be consistently excellent, varying on a range between 96% and 100% for both single and multiple stems predictions (Supplementary Table 1).

We tested whether MiRfold's two criteria (Billoud *et al.*, 2005), which are close to our Criteria III, IV and V, could provide good parameters to discriminate putative pre-miRNAs or not. MiRfold takes an miRNA sequence and looks for two extensions, downstream and upstream of the miRNA, having an optimal folding. Among all different possible extensions, MiRfold selects the one that minimizes a penalty score on miRNA/miRNA* matching and the MFE of the secondary structure. The first condition is an analog of our combinatorial Criterion III, where we include a dependence on the miRNA length though. The second condition is an analogue of our physico-chemical Criteria IV and V where we take into account the pre-miRNA length (IV) and the pre-miRNA GC-composition (V). We computed ROC curves by varying the two MiRfold's criteria on the multiple stems positive and negative datasets constructed on *H.sapiens*, *O.sativa*, *C.elegans*, *A.thaliana* and *R.norvegicus*. Results obtained with MIReNA and with MiRfold criteria are very different (Supplementary Fig. 7). MIReNA ROC curves upper bound ROC curves on MiRfold criteria in all datasets, and show prediction stability over a large parameter range (this corresponds to the overlapping of ROC curves) contrary to ROC curves corresponding to MiRfold criteria. We conclude that MiRfold criteria are not optimal for pre-miRNA discrimination. Notice that no use of MiRfold software was made but that only the two MiRfold criteria were tested.

Table 1. Comparison of MIReNA with different predictive systems on *H.sapiens* datasets

	MiPred dataset for *H. sapiens*			
Methods	Spe	Sen	Acc	MCC
MiPred	93.21	89.35	91.29	0.83
microPred	70.19	**92.39**	81.25	0.64
miR-abela	**99.25**	72.62	85.98	0.75
MIReNA	92.83	91.63	**92.23**	**0.84**

Single stem and multiple stems datasets for *Homo sapiens*

	Single stem				Multiple stems			
Methods	Spe	Sen	Acc	MCC	Spe	Sen	Acc	MCC
MiPred	96.21	86.36	**91.29**	**0.83**	N/A	N/A	N/A	N/A
microPred	73.33	**91.52**	82.42	0.66	78.94	**90.52**	84.73	0.70
miR-abela	**99.70**	65.45	82.58	0.69	**99.86**	62.06	80.96	0.67
MIReNA	92.27	90.30	**91.29**	**0.83**	94.42	90.10	**92.26**	**0.84**

Top: MiPred dataset (Jiang *et al.*, 2007); results reported for MIReNA correspond to best MCC in the ROC curve analysis for *H.sapiens* (Supplementary Fig. 6). Bottom: Predictions realized on datasets of secondary structures with single (left) and multiple (right) stems. Results reported for MIReNA correspond to best MCC in ROC curves analysis obtained by varying Criteria I–V (see Supplementary Table 1). Bold values are the highest in a column.

3.2 Comparison with other predictive tools

MIReNA has been compared to MiPred, microPred and miR-abela on two distinct datasets. The first dataset is the testing dataset used in Jiang *et al.* (2007) from *H.sapiens*. It is composed by a positive and a negative set of pre-miRNAs and pseudo pre-miRNAs where miRNAs positions are not specified. For comparison, we do not use any information coming from known miRNAs in miRBase, but merely test MIReNA on the brown path in Figure 1. The second dataset is composed by five negative and positive sets from five different species (including sets for *H.sapiens* that are different from the ones defining the first dataset). Positive sets have been constructed from species-specific precursors in miRBase and the negative ones from CDS sequences coming from the corresponding genomes.

3.2.1 Comparison with MiPred MiPred is a classifier that discriminates real single stem pre-miRNAs from pseudo ones (Jiang *et al.*, 2007). It is based on a Random Forest machine learning method using more than 30 pre-miRNA features. On the first dataset, MiPred obtains better specificity results, while best sensitivity, Acc and MCC are obtained by MIReNA (Table 1, top). On the

Table 2. MIReNA and miRDeep predictions on deep sequencing data for *C.elegans* and *H.sapiens*

Species	*Caenorhabditis elegans* and *Homo sapiens* deep sequencing datasets						
	Method	#pred precs	Sen	Signal to noise	#specif in miRBase	#new precs	
						All	Specif
Homo sapiens	miRDeep	284	70.55	8 : 1	31	64	30
	MIReNA	266	64.42	9 : 1	11	63	29
Caenorhabditis elegans	miRDeep	120	85.51	12 : 1	10	1	0
	MIReNA	116	79.71	17 : 1	2	5	4

MIReNA is used to predict multiple stems miRNA precursors using deep sequencing reads as potential miRNAs. We report: number of predicted precursors (3rd column), sensitivity (4th), signal-to-noise ratio (5th), number of specific (that is, captured by one method but missed by the other) miRNAs in miRbase version 14.0 (6th), total number of new predicted precursors (7th) and number of new specific predicted precursors (8th). An exact match with the miRNA in miRbase or with a read is required for the results in the last three columns. Results were obtained by using thresholds on Criteria I–V computed on the *C.elegans* and on the *H.sapiens* ROC curves (for multiple stems) displaying best MCC (Fig. 2d and Supplementary Fig. 3p).

second dataset, MIReNA performs slightly better than MiPred for all species when looking at Acc and MCC values (Table 1, bottom and Supplementary Table 2, computed by varying Criteria I–V). In conclusion, MIReNA obtains as good results as MiPred, or better, and simplifies in an essential manner the searching space, reducing it to 5 parameters instead of 30. It can be applied to larger sets of pre-miRNAs possibly presenting multiple stems.

3.2.2 Comparison with microPred microPred is a classifier system used to discriminate pre-miRNAs from other non-coding RNAs by using machine learning techniques (Batuwita and Palade, 2009). The comparison between microPred and MIReNA has been made on single and multiple stems datasets (Table 1 and Supplementary Table 2). MIReNA obtains a slightly lower sensitivity but much higher specificity, Acc and MCC values than microPred for four species. For *A.thaliana*, MIReNA outperforms microPred.

3.2.3 Comparison with miR-abela miR-abela identifies clustered miRNAs by searching for pre-miRNAs (Sewer *et al.*, 2005). It uses a machine learning method to computationally validate the pre-miRNAs. Sensitivity, Acc and MCC are higher for MIReNA on both testing datasets (Table 1 and Supplementary Table 2), with a lower specificity, making MIReNA a better classifier to discriminate pre-miRNAs from pseudo ones.

3.3 miRNA discovery from deep sequencing data

MIReNA can be applied to deep sequencing datasets (red path in Fig. 1). It has been compared to miRDeep on *C.elegans* and *H.sapiens* deep sequencing data (Table 2). MIReNA provides a slightly lower number of predictions with a lower sensitivity against a higher signal-to-noise ratio than miRDeep. The two tools appear complementary due to an important set of specific predictions revealed on the analysis of both *C.elegans* and *H.sapiens* data. This means that by running MIReNA one can recover a number of miRNAs in miRBase, which have been missed by miRDeep and viceversa. MIReNA predicts 5 (4 of which are MIReNA specific) pre-miRNAs that do not contain known miRNAs in miRBase for *C.elegans* and 63 (29 of which are MIReNA specific) for *H.sapiens*; miRDeep predicts 1 (non-specific) pre-miRNA that does not contain known miRNAs in miRBase for *C.elegans* and 64 (30 of which are miRDeep specific) for *H.sapiens*.

Deep sequencing data are used twice in the red path of Figure 1. The first time to select clusters of reads providing precursors as done in Friedländer *et al.* (2008). The second time when filtering pre-miRNA/miRNA pairs that respect the Dicer processing. Intuitively, given a pre-miRNA/miRNA pair, we expect reads to overlap the miRNA, the miRNA* and the remaining hairpin loop (i.e. the sequence between the end of the miRNA and the beginning of the miRNA*) without having too many reads overlapping two of these regions. This idea is present in Friedländer *et al.* (2008).

A main difference between MIReNA and miRDeep is the selection of the potential miRNA at the beginning of the algorithm (second box in red path of Fig. 1) that for MIReNA is less restrictive. We accept pre-miRNA/miRNA pairs where each miRNA matches a read in the dataset, while miRDeep only considers those pairs where the miRNA is most represented by reads matching the pre-miRNA. This means that for the same pre-miRNA, MIReNA may consider several pre-miRNA/miRNA pairs satisfying Criteria I–V, whereas miRDeep considers only one pair. However, notice that our Criteria I–III (depending on the length of the miRNA) turn out to be more restrictive than those used by miRDeep (based on fixed thresholds) to accept a miRNA in a pre-miRNA.

Another main difference between MIReNA and miRDeep is in the second filtering step of MIReNA (fourth box in red path of Fig. 1). MIReNA considers the same ideas used in miRDeep but encodes them in a set of combinatorial rules instead of defining a probabilistic model.

3.4 Predictive miRNA analysis with MIReNA

MIReNA has been applied to discover miRNAs and pre-miRNAs in species already known to contain miRNAs and in species where their existence could be verified experimentally after prediction. An exploratory analysis was done on six eukaryotic species. Two of them, *A.thaliana* and *C.elegans*, are multicellular and contain experimentally known miRNAs. The remaining four, the two diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, and the two yeasts *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*, are unicellular organisms not known to contain any miRNA.

Based on miRBase, a dataset of known miRNAs from multicellular species, MIReNA predicts similar miRNA sequences whose associated pre-miRNA structures satisfy necessary structural conditions and whose pre-miRNA sequences are required not to have low complexity (by RepeatMasker). Secondary structures of predicted pre-miRNAs might display multiple stems (Supplementary Fig. 1c and d).

MIReNA finds potential miRNAs for all six studied organisms (Table 3, 3rd column). When considering EST data, all species contain potential miRNAs whose pre-miRNA sequences match some EST (Boguski *et al.*, 1993) with the exception of yeast species having no EST match lying in intergenic regions (Table 3). Among the 1842 (respectively 621) predicted pre-miRNA/miRNA pairs of *C.elegans* (respectively *A.thaliana*), 1392 (respectively 370) are similar in sequence to known miRNAs whose hairpin sequences

Table 3. MIReNA's analysis of several genomes based on miRNAs from miRBase

Pre-miRNA/miRNA pairs predicted by MIReNA starting from miRNAs in miRBase

Species	Filter	Repbase	EST		Deep seq	
		Filter with Repeat Masker	#matching in dbEST		#matching of pred. miRNAs	
			Filter	New	Filter	New
Thalassiosira pseudonana (0)	88	74	10	10	N/A	N/A
Phaeodactylum tricornutum (0)	66	54	11	11	N/A	N/A
Schizosaccharomyces pombe (0)	18	12	0	0	N/A	N/A
Saccharomyces cerevisiae (0)	32	8	0	0	N/A	N/A
Caenorhabditis elegans (174)	3078	1842	54	54	424	343
Arabidopsis thaliana (190)	1124	621	100	84	360	243

MIReNA identification of pre-miRNA/miRNA pairs with miRNA sequences similar (but not necessarily identical) to miRNAs in miRBase. The number of known pre-miRNAs belonging to miRBase is given after the species name. The number of pre-miRNA/miRNA pairs that do not overlap CDS, scRNA, snRNA, snoRNA, tRNA, rRNA and 21U-RNA sequences on either strands is reported (2nd column; all columns named 'filter' identify a screening based on genomic information). For *T. pseudonana* and *P. tricornutum*, we check overlapping with CDSs (the only information available). A second filter of all pre-miRNA/miRNA pairs is realized with RepeatMasker (3rd), EST sequences (4th) and deep sequencing data (6th) and validation of putative pre-miRNAs is done by matching an EST to the pre-miRNA or a read to the miRNA. A number of new pre-miRNA/miRNA pairs, not contained in miRBase, are predicted from ESTs or deep sequencing data after filtering ('new').

are not classified within the family classification of miRBase. Of these 1392 (respectively 370) sequences, 718 (respectively 119) are similar to already known miRNAs from *C.elegans* (respectively *A.thaliana*). This means that there exists a large number of miRNA paralogs whose hairpin sequences have not been yet classified within miRBase.

Multicellular species contain potential miRNAs that match deep sequencing reads. MIReNA highlights that *new* expressed pre-miRNAs may exist (Table 3, 5th and 7th columns). The associated miRNAs are similar and not necessarily identical to known miRNAs in miRBase (possibly identified for other species). This pool of pre-miRNA/miRNA constitutes a set of new predictions, since the miRNA is not a known miRNA for the species.

4 METHODS

4.1 Genomic data and datasets

Genomes flat files for *A.thaliana*, *S.cerevisiae*, *S.pombe*, *H.sapiens*, *R.norvegicus* and *O.sativa* were retrieved from the NCBI site http://www.ncbi.nlm.nih.gov/genomes/, for *C.elegans* from the WormBase site http://www.wormbase.org, for *P.tricornutum* and *T.pseudonana* from the JGI site http://genome.jgi-psf.org. The corresponding assembling versions are listed in Supplementary Table 3. Known miRNA sequences and their genome location were downloaded from version 14.0 of miRBase (http://microrna.sanger.ac.uk/sequences/ftp.shtml).

miRBase (Griffiths-Jones, 2004; Griffiths-Jones *et al.*, 2006) was used as a reference dataset to define thresholds and to construct testing sets of pre-miRNA/miRNA pairs. It declares a miRNA to be experimentally validated or computationally predicted by coupling it with a specific pre-miRNA. In miRBase, there are miRNAs within a species that have been validated for several different pre-miRNAs, and there are several pre-miRNAs that validate several miRNAs contained in their sequence. Out of this list of 11 488 miRNA/pre-miRNA pairs, 6166 concern experimentally validated miRNAs and 5322 concern computationally predicted miRNAs. Among the 6166 experimentally validated pairs, 5471 present different pre-miRNA sequences; among the 5322 computationally predicted pairs, 4406 have different pre-miRNA sequences.

Thresholds in MIReNA were computed using experimentally validated pre-miRNA/miRNA pairs in miRBase only.

We constructed several datasets on the five species *H.sapiens*, *A.thaliana*, *C.elegans*, *O.sativa* and *R.norvegicus* following the generation method proposed in Xue *et al.* (2005). The five datasets have been used to compare MIReNA to available tools (Table 1 and Supplementary Table 2). Each dataset is constituted by two sets of sequences, a positive one, made of experimentally validated and computationally predicted pre-miRNA sequences in miRBase, and a negative one, made from coding sequences, that is extracted from concatenated CDS of the corresponding species. Positive and negative sets are composed of pre-miRNAs and pre-miRNAs-like sequences, i.e. sequences validating four conditions on: (i) sequence length; (ii) minimum number of basis paired on the stem of the hairpin structure; (iii) MFE of the sequence structure; and (iv) hairpin structure, made of a single stem. These four conditions have been previously used in MiPred. Sets contain from a minimum of 71 to a maximum of 660 hairpin structures characterized by a single stem. Positive and negative datasets are referred to as 'single stem datasets' in the text. We were interested to study hairpin structures with several stems too and generated for each species negative datasets accepting hairpins, following the same generation approach as in Xue *et al.* (2005). Corresponding positive datasets contain all known pre-miRNAs (from 173 to 717), for the considered species, in miRBase, and no filtering with conditions (i–iv) above has been applied. We refer to these positive and negative datasets as 'multiple stems datasets'.

The testing dataset described in Jiang *et al.* (2007) was used to compute Table 1 (top): 263 positive sequences are human miRNAs and 265 negatives are issued from CDS sequences as described above.

4.2 ROC curves

ROC curves are plot on datasets from five species by varying values of MIReNA parameters and MiRfold criteria. MIReNA's pre-miRNAs validations from the datasets were computed using the five criteria introduced above. As no potential miRNA is determined for pre-miRNAs in the datasets, we say that a pre-miRNA validates Criteria I, II and III if it includes a subsequence (i.e. a potential miRNA) of 18–25 nt in length that ensures the corresponding thresholds. MIReNA's ROC curves were computed for different combinations of Criteria I–V. Also, ROC curves have been plot based on MiRfold two criteria: a *penalty* score counting the number of mismatches between miRNA and miRNA*, and the MFE.

4.3 Computational tools used for comparison

MIReNA was compared to different existing tools. MiPred (Jiang *et al.*, 2007) implementation was provided to us by the authors (personal communication). We used miR-abela (with default parameters) at http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi, microPred at http://www.comlab.ox.ac.uk/microPred/microPred-server.html. miRDeep was downloaded at http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/

4.4 Computational tools used for MIReNA

The computation of secondary structures is realized with the RNAfold program of the Vienna RNA package (Hofacher *et al.*, 1994) by minimizing the free energy of the structure. We modified the source code of RNAfold to avoid the sequential computation of all possible subsequences, and directly compute the secondary structure of the whole sequence and retrieve the minimum free energy structure of all sub-sequences of it by backtracking.

This adaption of the code concerns Step 4 in the algorithm (Fig. 1). RNAfold is found at http://www.tbi.univie.ac.at/RNA/. Default parameters of RNAfold were used.

RepeatMasker [A.F.A.Smit *et al*. (2009) RepeatMasker unpublished data] has been used as a filter to remove sequences containing repeats. We used version open-3.2.8 with RMLib: 20090604 (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) and *crossmatch* as search engine with a *slow* speed. Repeat sequences used by RepeatMasker are stored in the Repbase database (Jurka *et al.*, 2005).

4.5 Deep sequencing data and their mapping

The two sets of reads used to compare MIReNA and miRDeep are constructed following Friedländer *et al.* (2008). The first puts together two 454 deep sequencing *C.elegans* datasets of GEO database at NCBI, and the second puts together two Solexa and one 454 deep sequencing *H.sapiens* datasets of GEO database at NCBI (see Supplementary Table 4 for accession numbers). Putative pre-miRNAs were constructed by using scripts from Friedländer *et al.* (2008) realizing the excision of clusters of reads. To compute sensitivity of the two systems, we considered miRNAs in miRBase version 14.0. This version of miRBase contains 13 *C.elegans* and 9 *H.sapiens* miRNAs predicted by miRDeep in Friedländer *et al.* (2008), which were not contained in miRBase version 10.0 used in Friedländer *et al.* (2008). Sensitivity is computed as the ratio between the number of species-specific miRNAs of miRBase contained in predicted precursors and the number of species-specific miRNAs in miRBase matching deep sequencing reads (a match allows for mismatching in the last three nucleotides of miRNAs). Specificity cannot be directly computed because of the impossibility to directly evaluate false positives. This value can be estimated, following Friedländer *et al.* (2008), by generating a set of pre-miRNAs together with positions of mapped reads. To do so, we start from the set of putative pre-miRNAs where positions of mapped reads are identified and shuffle the association between the set of pre-miRNAs and the set of configurations of read positions. We obtain a set of negative pre-miRNAs on which the algorithm is tested and the number of false positives is computed. This is done 100 times and a distribution of the number of false positives is considered. The signal-to-noise ratio is $N/\mu_{\mathrm{FP}}:1$ where μ_{FP} is the mean of the distribution and N the number of predictions obtained with MIReNA or miRDeep.

The mapping of a read is done asking for 100% of identity over the full length of the read, for all deep sequencing datasets, using megablast (with options `-W12 -p100`). A read is considered only when it perfectly maps ≤ 5 locations in the genome. The only exception is done with the *H.sapiens* dataset GSE10829 due to our comparison with miRDeep: given a read, a mapping of the read to possibly several sites in the genome, is constructed by asking that the first 18 nucleotides of the read perfectly align with the genomic sequence and by extending the alignment on the 3′-end until the first mismatch is found. Accepted matches are those with maximal length. There might be several of them and reads with more than five matches are discarded.

The red path in the algorithm asks for a filtering step for precursors that is based on a Dicer processing condition. Following Friedländer *et al.* (2008), we define a Dicer processing signature for a pre-miRNA/miRNA pair to be as follows: a read has to map with the potential miRNA of the pair, or its corresponding miRNA* or the remaining hairpin loop. The alignment can extend the miRNA or the miRNA* on the 5′-end by at most 2 nt and the 3′-end by at most 5 nt. Among the reads mapping the precursor, we ask at least the 90% of them to satisfy the signature.

Pre-miRNA/miRNA pairs satisfying the Dicer processing condition should also satisfy at least one of the following conditions: (i) the miRNA* is mapped by a read where the alignment can extend the miRNA* on the 5′-end by at most 1 nt and the 3′-end by at most 5 nt; (ii) the miRNA sequence should be matched by at least two reads (where the alignment can extend the miRNA on the 5′-end by at most 2 nt and the 3′-end by at most 5 nt) and within nucleotide 2–8 from the 5′-end should match nucleotides 2–8 of a known miRNA in miRBase (miRNAs for metazoan are considered with the exception of the species under consideration); (iii) the different pre-miRNA subsequences mapped by reads that satisfy the Dicer processing condition should be at least 3 (4) for *C.elegans* (*H.sapiens*). The threshold 3 (4) has been statistically evaluated by mapping reads on the *C.elegans* (*H.sapiens*) pre-miRNAs in miRBase, by retaining only those pre-miRNAs that do not satisfy conditions (i) and (ii), and by computing the mean of the number of different pre-miRNA subsequences mapped by reads that satisfy the Dicer processing condition. Conditions (i–iii) are inspired by the criteria used in Friedländer *et al.* (2008) to define a probabilistic scoring.

In Table 3, to filter MIReNA miRNA predictions we used deep sequencing data (see Supplementary Table 4 for GEO accession numbers).

4.6 Pre-miRNAs filtering using ESTs or reads

In columns 4–5 of Table 3, we filtered MIReNA predicted pre-miRNA sequences against the EST database dbEST of the studied species (Boguski *et al.*, 1993) by using BLAST (http://blast.ncbi.nlm.nih.gov/) with default parameters: a pre-miRNA sequence matches an EST when at least 80% of similarity on at least 80% of the sequence with an *e*-value $<=1e^{-30}$ are present. In columns 6–7 of Table 3, miRNA predictions from deep sequencing data are obtained starting from pre-miRNA/miRNA pairs where the miRNA is similar to a known miRNA in miRBase, and by filtering out those miRNAs that do not match a read. For the matching of the read, we ask for 100% identity on the miRNA length.

4.7 Prediction system assessment

To properly evaluate MIReNA performance, we rely on the following quantities: the number of known pre-miRNAs correctly predicted (true positives, TP), the number of pseudo pre-miRNAs correctly predicted (true negatives, TN), the number of pseudo pre-miRNAs incorrectly predicted as pre-miRNAs (false positives, FP) and the number of known pre-miRNAs incorrectly predicted as pseudo pre-miRNAs (false negatives, FN). We use four standard measures of performance: Specificity ($\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN+FP}}$), Sensitivity ($\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP+FN}}$), Acc ($\mathrm{Acc}=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$) and MCC ($\mathrm{MCC}=\frac{\mathrm{TP\times TN-FP\times FN}}{\sqrt{(\mathrm{TP+FP})\times(\mathrm{TN+FN})\times(\mathrm{TP+FN})\times(\mathrm{TN+FP})}}$) (Matthews, 1975).

4.8 Implementation

MIReNA package is written in bash, C, perl, python, and it is available at http://www.ihes.fr/~carbone/data8/. Parameters are described in Supplementary Material and in the help file by typing ‘./MIReNA.sh -h’.

5 DISCUSSION

MIReNA is a platform that helps to test new hypothesis on miRNAs and pre-miRNA structures. Across organisms, it might be that the characteristic length of miRNA sequences might vary slightly and a tool like MIReNA, allowing for length calibration, can help *in silico* investigation of the existence of miRNA sequences with non-standard properties. Similar considerations might be extended to unusual characteristics of pre-miRNA structures as hairpin multiple branching, contrary to what has been usually assumed (Lim *et al.*, 2003a, b; Nelson *et al.*, 2003; Wang *et al.*, 2005). Experimentally validated multiple branched pre-miRNAs exist (Supplementary Fig. 1c) and in miRBase, they constitute the 10% (671 over 6166) of experimentally validated pre-miRNAs. Many other properties of miRNAs and pre-miRNAs can be fine-tuned in MIReNA giving the opportunity to the biologist to perform an *in silico* identification of miRNA sequences (possibly with non-standard features) in experimentally unexplored genomes.

MIReNA does not search for structurally similar pre-miRNAs. In consequence, its results are different from those obtained by most

methods (Dezulian *et al.*, 2006; Hertel *et al.*, 2006; Legendre *et al.*, 2005), and in particular from those based on secondary structure conservation (Lim *et al.*, 2003a; Wang *et al.*, 2005). One should expect MIReNA to detect pre-miRNAs that cannot be predicted by other approaches. (See miRscan; Lim *et al.*, 2003a, analysis in Supplementary Material.)

MIReNA handles experimental data and can check the compatibility of its predictions against EST data and deep sequencing data. Deep sequencing data can be used in two distinct manners. MIReNA can check either that predictions from miRNAs in miRBase are matched by reads, or it can directly employ reads to predict pre-miRNA/miRNA pairs, as previously done in miRDeep. MIReNA and miRDeep provide a large core of common predictions but also specific ones. MIReNA appears to have a signal-to-noise ratio higher than miRDeep and its specific predictions highly support experimental validation.

The exploratory analysis done on six eukaryotic species highlights the existence of pre-miRNA/miRNA pairs for all species when a purely genomic prediction filtered with RepeatMasker is realized (Table 3, 3rd column). On the other hand, when ESTs are taken into account, miRNAs are predicted only for multicellular species and for diatoms (Table 3, 4th column). The number of pre-miRNA/miRNA pairs is much smaller for unicellular species than for multicellular ones, and this difference is consistent with the observed correlation between expansion of sets of miRNAs and morphological complexity of the organisms (Heimberg *et al.*, 2008; Lee *et al.*, 2007; Niwa and Slack, 2007; Sempere *et al.*, 2006; Wheeler *et al.*, 2009).

The few miRNAs found for diatoms suggest diatoms miRNAs to have specific characteristics and be different from those observed up to now, their detection demanding different parameterizations. The existence of RNA interference (RNAi) in diatoms has been indirectly proven by using a putative silencing machinery; several enzymes have been proposed to participate in silencing but the identification of miRNAs is still undergoing (De Riso *et al.*, 2009).

MIReNA found no pre-miRNA in yeasts (after filtering with EST data) in agreement with the absence of RNAi machinery. miRNAs have never been reported for fungi and are absent in *S.pombe*, which has known RNA silencing mechanisms (Cerutti and Casa-Mollano, 2006; Hertel *et al.*, 2006). The gene silencing pathway of RNAi was estimated to have recently being lost in *S.cerevisiae*, where reintroducing Dicer and argonaute restores RNAi (Drinnenberg *et al.*, 2009). The finding of eight potential miRNAs in *S.cerevisiae* from genomic analysis (Table 3, 3rd column) is not in contradiction with the loss of the RNAi machinery but might suggest the existence of an ongoing process of hairpin deletion.

EST sequences and deep sequencing data depend on specific experimental conditions and pre-miRNAs predictions filtered or computed from these data capture only partially miRNAs expressed in the cell. Their exploration allows us to predict new miRNAs (Tables 2 and 3) though and parameterizable tools like MIReNA open a way to miRNA discovery in organisms where miRNA characteristics are yet unknown.

MIReNA answers to several questions addressed by available tools, it performs at least as well as these tools and in most cases better than those, displaying very good sensitivity, specificity and Acc. The biologist can use it as a platform to check several questions around miRNAs and pre-miRNAs prediction starting from different kinds of available information (computationally and experimentally validated miRNAs, or deep sequencing data) and can play with sensitivity and specificity by using parameters variability.

ACKNOWLEDGEMENTS

Part of our computations were realized on the computers of the Centre de Calcul Recherche (CCR), Université Pierre et Marie Curie.

Funding: Ministère de l'Enseignement Supérieure et de la Recherche (MESR) doctoral fellowship and a MESR teaching assistantship (ATER) (to A.M.).

Conflict of Interest: none declared.

REFERENCES

Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.

Billoud,B. *et al.* (2005) Identification of new small non-coding RNAs from tobacco and Arabidopsis. *Biochimie*, **87**, 905–910.

Boguski,M.S. *et al.* (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.

Cerutti,H. and Casas-Mollano,J.A. (2006) On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.*, **50**, 81–99.

Delisi,C. and Crothers,D.M. (1971) Prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **68**, 2682–2685.

De Riso,V. *et al.* (2009) Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.*, **37**, e96.

Dezulian,T. *et al.* (2006) Identification of plant microRNA homologs. *Bioinformatics*, **22**, 359–360.

Drinnenberg,I.A. *et al.* (2009) RNAi in budding yeast. *Science*. **326**, 544–550.

Friedländer,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Griffiths-Jones,S. (2004) The miRNA Registry. *Nucleic Acids Res.*, 2004, **32**, D109–D111.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Heimberg,A.M. *et al.* (2008) MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA*, **105**, 2946–2950.

Hertel,J. *et al.* (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 15.

Hertel,J. and Stadler,P.F. (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197–e202.

Hofacker,I.L. *et al.* (1994) Fast Folding and Comparison of RNA Secondary Structure. *Monatsh. Chem.*, **125**, 167–168.

Jiang,P. *et al.* (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339–W344.

Jurka,J. *et al.* (2005) Repbase update, a database of eukaryotic repetitive elements. *Cyrogenet. Genome Res.*, **110**, 462–467.

Lai,E.C. *et al.* (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.

Lee,R.C. *et al.* (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lee,C.T. *et al.* (2007) Evolutionary conservation of microRNA regulatory cicuits: an examination of microRNA gene complexity and conserved microRNA-target interactions. *DNA Cell Biol.*, **26**, 209–218.

Legendre,M. *et al.* (2005) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **21**, 841–845.

Levenshtein,V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Doklady*, **10**, 707–710.

Lim,L.P. *et al.* (2003a) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.

Lim,L.P. *et al.* (2003b) Vertebrate microRNA genes. *Science*, **299**, 1540.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Myers,G. (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, **46**, 395–415.

Nelson,P. *et al.* (2003) The microRNA world: small is mighty. *Trends Biochem. Sci.*, **28**, 534–540.
Niwa,R. and Slack,F.J. (2007) The evolution of animal microRNA function. *Curr. Opin. Genet. Dev.*, **17**, 145–150.
Pasquinelli,A.E. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.
Sempere,L.F. *et al.* (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J. Exp. Zool.*, **306B**, 575–588.
Sewer,A. *et al.* (2005) Identification of clustered microRNA using an *ab initio* prediction method. *BMC Bioinformatics*, **6**, 267.
Tanzer,A. *et al.* (2008) Evolutionary genomics of microRNAs and their relatives. In Caetano-Anolles,G. (ed.) *Evolutionary genomics*. Wiley-Blackwell, Hoboken, 2010.
Tinoco,I. *et al.* (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
Wang,X. *et al.* (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.
Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.
Wheeler,B.M. *et al.* (2009) The deep evolution of metazoan microRNAs. *Evolution & Development*, **11**, 50–68.
Xu,Y. *et al.* (2008) MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, **24**, i50–i58.
Xue,C. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *Bioinformatics*, **6**, 310.
Zhang,Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203.
Zhang,B.M. *et al.* (2006) Evidence that miRNAs are different from others RNAs. *Cell. Mol. Life Sci.*, **63**, 246–254.