W68–W76 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 11 May 2009*
doi:10.1093/nar/gkp347

miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments

Michael Hackenberg[1], Martin Sturm[2], David Langenberger[3,4], Juan Manuel Falcón-Pérez[5] and Ana M. Aransay[1,*]

[1]Functional Genomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain, [2]Institute for Bioinformatics and Systems Biology, German Research Center for Environmental Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, [3]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universitat München, 85350 Freising, [4]Bioinformatics Group, Department of Computer Science, University of Leipzig, Haertelstr. 16-18, D-04107 Leipzig, Germany and [5]Metabolomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain

Received February 28, 2009; Revised April 13, 2009; Accepted April 22, 2009

ABSTRACT

Next-generation sequencing allows now the sequencing of small RNA molecules and the estimation of their expression levels. Consequently, there will be a high demand of bioinformatics tools to cope with the several gigabytes of sequence data generated in each single deep-sequencing experiment. Given this scene, we developed miRanalyzer, a web server tool for the analysis of deep-sequencing experiments for small RNAs. The web server tool requires a simple input file containing a list of unique reads and its copy numbers (expression levels). Using these data, miRanalyzer (i) detects all known microRNA sequences annotated in miRBase, (ii) finds all perfect matches against other libraries of transcribed sequences and (iii) predicts new microRNAs. The prediction of new microRNAs is an especially important point as there are many species with very few known microRNAs. Therefore, we implemented a highly accurate machine learning algorithm for the prediction of new microRNAs that reaches AUC values of 97.9% and recall values of up to 75% on unseen data. The web tool summarizes all the described steps in a single output page, which provides a comprehensive overview of the analysis, adding links to more detailed output pages for each analysis module. miRanalyzer is available at http://web.bioinformatics.cicbiogune.es/microRNA/.

INTRODUCTION

The recent years witnessed a profound change in our understanding of the regulation of gene expression. Small non-coding RNA especially came into focus as it became clear that they are key players in many cellular processes by post-transcriptionally regulating gene expression via either degradation, translational repression, or both (1,2). MicroRNAs, belonging to the family of small non-coding RNAs, are endogenous in many animal and plant genomes and are now recognized to be one of the major regulatory gene families in eukaryotic cells. They are believed to regulate the expression of around one third of all genes in the human genome, involved in many fundamental processes like metabolism, development and regulation of the nervous and immune systems (3,4). Furthermore, it has been reported that some microRNAs are actively involved in the development of pathologies like cancer (5).

The traditional experimental approach to measure the expression levels of microRNAs involves cloning and Sanger sequencing. This is an expensive and time-consuming procedure, and as a consequence, relatively little expression data is currently available [see (6) for a microRNA expression atlas]. Moreover, the huge range of microRNA expression from tens of thousands to just few molecules per cell complicates the detection of microRNAs expressed at low copy numbers. Hence many undetected microRNA may exist even in well-explored species. Recently, microRNA expression profiling panels became available for measuring expression levels by means of hybridization. These panels allow a

*To whom correspondence should be addressed. Tel: +34 944 061 325; Fax: +34 944 061 324; Email: amaransay@cicbiogune.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

high-throughput detection of microRNA expression. However, they do not allow the detection of new microRNAs.

Next generation sequencing platforms like Genome Analyzer (Illumina Inc.) or Genome Sequencer™ FLX (454 Life Science™ and Roche Applied Science) became recently available for the sequencing of small RNA molecules, which allows both the detection of expression levels and new microRNA sequences at high speed and sensitivity and low cost. However, each sequencing experiment produces up to 3 Gbp of sequence data, whose analysis represents an important bioinformatics challenge.

Given the importance of microRNAs in the regulation of gene expression, in the coming years many deep-sequencing experiments will be carried out to detect and measure their expression. Therefore, user friendly tools are required for the processing of the enormous amount of data that will be generated. To our knowledge, so far there is only one standalone tool available for the complete analysis of deep sequencing microRNA data: miRDeep published by Friedländer *et al*. (7). Specific software for SOLiD data does exist that allows the detection of known microRNAs but not the prediction of new microRNAs (http://solidsoftwaretools.com/gf/project/rna2map).

On the other hand, the prediction of microRNA genes has been extensively employed over the past years and several distinct approaches have been developed. Some of the methods used in the purely computational detection approaches were, for example, conservation of certain regions—phylogenetic shadowing (8), different machine learning methods like support vector machines using structure-sequence features (9), random forest models (10) or probabilistic co-learning models (11). Bentwich *et al*. (12) used further features like the stability of the hairpin together with an experimental validation. The main drawbacks of these approaches are that they are either limited to conserved microRNAs or that they tend to have a high rate of false positive predictions. However, new sequencing experiments open new possibilities in the prediction of microRNAs, allowing the generation of previously unavailable characteristics like, for example, the traces left by dicer processing.

Consequently, we have developed miRanalyzer, a web server tool which implements all necessary methods for a comprehensive analysis of deep-sequencing experiments of small RNA molecules. It detects known microRNAs annotated in miRBase and matches in other transcribed sequences (RNA, RFam and RepBase). Furthermore, miRanalyzer implements a highly accurate machine learning algorithm to predict new microRNAs (area under the curve—AUC—value of 97.9%). The algorithm is based on the *random forest* classifier and was trained on experimental data. This high accuracy is important for the identification of novel microRNAs, a process which usually results in high false positive rates. The tool also includes a Perl script for the proper generation of the input file using the Genome Analyzer (Illumina Inc.) pipeline results. Currently, miRanalyzer works for seven frequently used model species (human, mouse, rat, fruit-fly, round-worm, zebrafish and dog).

miRANALYZER

Input file description

A usual next-generation sequencing experiment produces up to several gigabytes of output corresponding to several hundred million base pairs. That is by far too many data to send over the web to analyze it using a web server tool. However, some reads (tags) obtained in microRNA sequencing experiments can be found multiple times in the output. The number of copies detected for a unique read is proportional to its expression level. Given this redundancy, the only information needed for the analysis of microRNAs are the sequences of the reads and the number of times each unique read was encountered in the experiment. This reduces the size of the input file drastically to a few megabytes, which is an acceptable size for a web server tool.

The tool accepts two different input formats (see http://web.bioinformatics.cicbiogune.es/microRNA/manual.html):

(i) a tab separated file with the read sequences and its counts (number of times each read has been obtained in the experiment) and
(ii) a multifasta file with the copy number of the unique reads (read count) as the description in the header (e.g. >ID 'count').

Along with this web-tool, we supply a Perl script, which counts the reads of a Genome Analyzer (Illumina Inc.) experiment, producing the tab separated input format. The script allows averaging of several lines, filtering for low quality reads and a simple analysis of differential expression ($\log_2$ ratios between different lines). A more detailed description of the Perl script can be found on the tutorial page (http://web.bioinformatics.cicbiogune.es/microRNA/manual.html).

Input parameters

Apart from the file with the read sequences, several other input options are available as summarized in Table 1. The parameters are explained in more detail in the corresponding sections of the manuscript and on the tutorial page (http://web.bioinformatics.cicbiogune.es/microRNA/manual.html).

WORKFLOW

miRanalyzer follows three internal analysis steps (Supplementary Figure S1): (i) detection of known microRNAs, (ii) mapping against libraries of transcribed sequences (mRNA, ncRNA, etc.) and (iii) prediction of new microRNAs. After each of these three steps, the detected reads are removed from the input data following the options set by the user (Table 1).

Detection of known microRNAs

In many of the microRNA experiments, the main purpose will be the detection of the expression levels of known microRNAs (or frequently the differential expression of microRNAs between two samples). Therefore, as the

Table 1. Summary of miRanalyzer input options

Input option	Description
Species	The species from which the input reads have been obtained
Number of mismatches	For the detection of known microRNAs the user can allow matches with up to two mismatches
Target gene method	Selection of the microRNA target gene prediction method for the ontological analysis.
Posterior probability threshold	The threshold for the posterior probability calculated by the classification model.
Considering adapter sequences	The read sequences frequently contain adapter sequences at its 3′ end. In this case, the user can take it into account by aligning also sub-sequences of a given minimum length (Data and methods section).
Detect just new microRNAs	This option skips the detection of known microRNAs.
Remove all mRNA matches	This option removes all reads which have been perfectly aligned with mRNA sequences. If this option is not set, the program will remove all reads which match in more than five mRNAs as we observed that these reads are frequently poly-A like sequences.
Remove RFam/RepBase.	These options remove all reads which have mapped to RFam or RepBase.
Just predict conserved microRNAs	This option limits the prediction of new microRNAs to regions which overlap with a Phylogenetically Conserved Element (PhastCons).

first analysis step, miRanalyzer detects the reads which correspond to known microRNAs. To carry out the detection of known microRNAs, we used the miRBase repository (13) which offers mature (the mature sequences of known microRNAs), mature-star (the sequence which pairs with the mature microRNA in the pre-microRNA secondary structure) and precursor microRNA sequences (sequence of the hairpin). For some of the microRNA precursors, it is unclear which of the two sequences (mature or mature-star) is biologically functional. In the case where both sequences are found to be expressed and the predominant product can be clearly detected, the minor product is labeled with a * (mature-star). Apart from the known mature-star sequences we generated a library with all other theoretically possible mature-star sequences. This also allows the detection of functional mature-star microRNAs whose expression has not been observed previously.

Many microRNA sequences, especially those belonging to the same microRNA family, exhibit a high degree of sequence similarity. Given that sometimes the read might be rather short (16 bp), non-unique matches might occur. A non-unique match exists if a read maps with the same quality (same number of mismatches) at different positions or to more than one sequence in the library. Often, alignment programs such as ELAND (included in Illumina Inc. pipeline) do not report these ambiguous matches. However, this might result in a loss of important information. Therefore, miRanalyzer reports these ambiguous matches, stating all microRNAs where matches have been found. Note that the groups of microRNAs that have been detected by the same read will normally belong to the same family.

The exact order of mapping against known microRNAs is: mature, mature-star, unknown mature-star and precursors/hairpin. Both unique matches (a read matches just to one known microRNA) and ambiguous matches (a read matches several microRNAs with the same quality) are detected and removed from the input at each step. The removal is important as otherwise the reads would be detected again in the precursor sequences (hairpins).

After known microRNAs detection, the corresponding target genes (those genes which are predicted to be regulated by the detected microRNA) are extracted (see 'Data and Methods' section) and pre-calculated ontological analyses are made available. In the case of ambiguous matches where the set of target genes is made up of a combination of various microRNAs, a link to Annotation-Module (14) is offered to launch the ontological analysis with the obtained gene list.

Mapping against transcribed sequences

After detecting reads that correspond to known microRNAs, miRanalyzer maps the remaining reads to databases of transcribed sequences as mRNA, non-coding RNA (RFam) and (retro)-transposons. Only perfect matches are considered in this analysis. These alignments are performed to achieve several aims:

- First, the mapping against the transcriptome should not yield any matches except for exonic microRNAs (1). Therefore, the number of matches can be viewed as a sample quality parameter (i.e. contamination of the RNA sample with degradation products and poly-A tails).
- Second, the mapping to RFam (and other libraries of ncRNA) and RepBase has two goals: (i) it might be interesting to see which other known small ncRNAs are in the sample and (ii) the removal of these reads will lower the number of false positives in the prediction of new microRNAs (small ncRNA might be confused with microRNAs). The removal of those sequences is optional (Table 1).
- Third, we also used the genomic annotation of repeats and transposons derived by RepeatMasker (http://www.repeatmasker.org). After aligning all reads with the genome, miRanalyzer checks if the read coordinates overlap with those of the RepeatMasker annotation. In this way we can detect reads that overlap with 'degraded' transposons whose expression might indicate 'domestication' (acquired function).

Predicting new microRNAs

The detection of new, previously unreported microRNAs is a very important analysis step in miRanalyzer tool as (i) a controversy exists over the real number of microRNAs (15) and therefore it is important to mine sequencing

experiments for new previously undetected microRNAs and (ii) for many species there are none or just a few microRNAs known. Consequently, the analysis of sequencing experiments in these species relies almost completely on the prediction of new microRNAs. Therefore, we set up a machine learning approach based on the random forest method (16) with a broad range of features. To train only on the most relevant features, we also employed a feature selection approach (see 'Data and Methods' section for detail).

We used three different data sets from human (hsa), rat (rno) and *Caenorhabditis elegans* (cel, see 'Data and Methods' section) for building the final prediction model. The results shown in Table 2 suggest that the classifier is highly sensitive and specific not only according to a standard 10-fold cross-validation, but also in a cross-species test on completely unseen test data. The results shown in the upper part of Table 2 depict the outcome when learning with one of the species (training set) and predicting the remaining ones (test data). For evaluation of prediction power in the same species, we applied a 10-fold cross validation approach. It can be seen that while the cross-validated results are high, the recall is moderate predicting on unseen data. We highlighted (yellow) the worst prediction values on the different test sets, which are 0.66 (cel/rno), 0.48 (rno/cel) and 0.64 (rno/hsa). To check whether we can improve prediction power for those in particular, we merged two datasets and evaluated against the third set (values highlighted in green). It can be seen that the prediction improved significantly, especially for *C. elegans*. While trained solely on rat or human and evaluated on worm a recall of only 0.48 and 0.67, respectively, could be reached. The merged training set, however, achieves a recall of 0.71, suggesting synergetic effects when integrating instances from different species into the training set. To benefit most from this effect, we trained the final classifier on all three data sets. Thus we obtain an area under the curve (AUC) value of 97.9% with a true positive rate of 0.79 and a false positive rate of 0.007 for the fixed threshold at 0.9. To test for robustness, we repeated the cross validation on 10 different negative sets, which resulted in a mean AUC value, true positive rate and false positive rate of 97.9%, 0.79 and 0.0077 with the standard deviations of 0.001, 0.01 and 0.003,

Table 2. The true positive rates (top part) and false positive rates (bottom part) for different classifiers at a posterior probability threshold of 0.9

Training set	Test set						
	rno	cel	hsa	rno-cel	rno-hsa	cel-hsa	rno-cel-hsa
True positive rate (threshold: 0.9)							
rno	0.74^{CV}	0.48	0.64	0.66	0.73	0.57	0.65
cel	0.66	0.77^{CV}	0.69	0.80	0.68	0.79	0.76
hsa	0.74	0.67	0.77^{CV}	0.70	0.84	0.81	0.79
rno-cel	0.89	0.91	0.75	0.79^{CV}	0.80	0.82	0.84
rno-hsa	0.91	0.71	0.93	0.80	0.78^{CV}	0.84	0.86
cel-hsa	0.74	0.91	0.91	0.83	0.84	0.81^{CV}	0.86
rno-cel-hsa	0.89	0.91	0.90	0.91	0.91	0.92	0.79^{CV}
False negative rate (threshold: 0.9)							
rno	0.01^{CV}	0.008	0.009	0.004	0.008	0.001	0.005
cel	0.005	0.004^{CV}	0.003	0.002	0.01	0	0.005
hsa	0.005	0.004	0.01^{CV}	0.01	0.01	0.005	0.005
rno-cel	0.02	0.008	0.01	0.009^{CV}	0.01	0.007	0.01
rno-hsa	0.02	0.01	0.01	0.01	0.01^{CV}	0.01	0.01
cel-hsa	0.005	0.004	0.009	0.004	0.01	0.003^{CV}	0.01
rno-cel-hsa	0.01	0.004	0.003	0.01	0.01	0.009	0.007^{CV}

The superscripted 'CV' denotes that this value was achieved in a standard 10-fold cross-validation approach. The highlighted false positive rates correspond to the true positive rates discussed in the text.

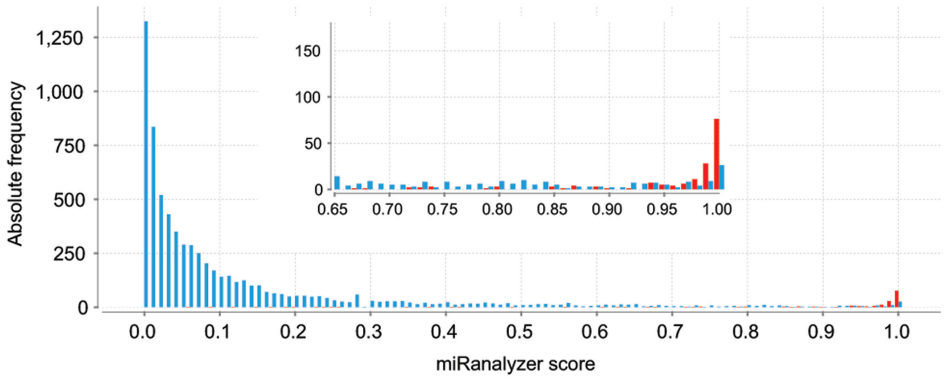

Figure 1. Histogram of miRanalyzer scores. Known microRNAs are colored in red, all other data are colored in blue. The insert is a close-up for candidates with scores better than 0.65.

respectively. Note that, the human and *C. elegans* sets where also used by Friedländer *et al.* who reported a recall of 89% on *C. elegans* and 72% on human. Table 2 shows that our approach reaches a recall of 75% on human when trained on rat/*C. elegans* (predicting on unseen data) and 91% on *C. elegans* using the final prediction model (predicting on previously seen data).

Figure 1 shows a cross-species evaluation of miRanalyzer trained on human and *C. elegans* and evaluated on rat. Obviously, most of the data have very low scores (the posterior probability assigned by the classification model to each instance) assigned. We build a close-up for the range between 0.65 and 1 to better visualize the high scoring predictions. It can be seen that the known rat microRNAs are strongly accumulated towards scores of 1, demonstrating the high predictive power of our approach and the good ability to generalize. Note that the classifier has never seen data from rat before. See also Supplementary Figures S3 and S4 for graphical representation of other quality parameters and Receiver Operating Characteristics of different classifiers discussed in this section.

A WORKING EXAMPLE

As a working example we used data derived from an experiment carried out in our laboratory with rat hepatocytes following standard protocols for smallRNA sample preparation and deep-sequencing (http://www.illumina.com/). Figure 2 shows the summary output page of miRanalyzer run on these data. The page is made up of five boxes that reveal the intrinsic workflow of miRanalyzer.

The first box shows the current state of the process (executing, pending, etc.) on the left side and depicts a short summary of the process (input data and options) on the right side.

The second box shows the summary of the analysis of known microRNAs. Each column corresponds to the mapping against a different set of sequences (mature, mature-star, etc.). The last row provides a link to detailed output for each of the columns. For example, the analysis of unknown mature-star sequences shows that miR-423-star is moderately expressed (744 copies) while the sequence which is annotated in miRBase (mature miR-423) has less than 10 copies (Supplementary Figure S2).

The third box summarizes the matching of reads to several sets of transcribed sequences. For example the fraction of reads mapped to the transcriptome may give a good estimate on the sample quality. It can be seen that around 8.3% of all reads in this sample originate from mRNA but this corresponds just to 3% of transcription amount (number of mRNA reads/total number of reads).

The fourth box shows the summary of the detection of new microRNAs. In addition, a link is given for further information on each read cluster that has been predicted to be a novel microRNA (Supplementary Figure S5). A link is also provided to a detailed output page with information on the chromosomal coordinates, the long hairpin structure and a verification if the reads have been detected before in the experiment (for example if matched against RepBase, etc.).

Finally, the last box gives a summary of the filtered and unmapped reads.

CONCLUSIONS

miRanalyzer is a web server tool for the integral analysis of next generation sequencing data of small RNA molecules. It allows both the detection of known microRNAs and the prediction of new microRNAs. For the prediction of new microRNAs a new sensitive machine learning algorithm was developed which reaches an AUC of 97.9% in our tests. Furthermore, the tool detects matches of the reads against other libraries of transcribed sequences such as mRNA, RFam (RNA) and RepBase (Transposons). Currently, the tool works for seven species, but can easily be extended upon request.

DATA AND METHODS

Sequence data

miRanalyzer uses the newest genome assembly of each available species which were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html): *Homo sapiens* (hg18, NCBI 36.1), *Mus musculus* (mm8, NCBI 36), *Rattus norvegicus* (rn4, version 3.4), *Drosophila melanogaster* (dm3, BDGP Release 5), *Caenorhabditis elegans* (ce6, WUSTL School of Medicine GSC and Sanger Institute version WS190), *Canis familiaris* (canFam2, v2.0) and Danio rerio (danRer5).

The mRNA sequence data were derived from different databases: *H. sapiens*, *M. musculus*, *R. norvegicus* and *D. rerio* from NCBI RefSeq (ftp://ftp.ncbi.nih.gov/refseq/), *D. melanogaster* from FlyBase (http://flybase.org/) and *C. elegans* from WormBase (http://www.wormbase.org/). The mRNA sequences for *C. familiaris* were extracted from the genomic sequence using the Galaxy platform (17).

In addition, mature microRNA sequences were derived from mirBase version 12.0 (http://microrna.sanger.ac.uk/sequences/); RNA sequences included in RFam version 9.0 (18) were downloaded from http://rfam.sanger.ac.uk/; and RepBase version 10.10 (19) were obtained from http://www.girinst.org/. Annotations and genomic coordinates of RepeatMasker and PhastCons elements where downloaded from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start).

We used deep–sequencing data from three different experiments: (i) the combined *C. elegans* data (accession no. GSE6282 and GSE5990 from GEO database at NCBI), which have been used also in (7) with a total of 205 575 unique reads, (ii) data from human HeLa cells (7) with accession no. GSE10829 and 319 939 unique reads and (iii) data from rat hepatocytes generated in our lab, available on our website (http://web.bioinformatics.cicbiogune.es/microRNA/defaultReads.txt) with 22 086 unique reads.

CIC bioGUNE

miRanalyser:

A microRNA detection and analysis tool for next-generation sequencing experiments

Restart | Tutorial & Test Data | Perl script

Queing and Execution

The program has finished

You can bookmark this page

Download all results in plain text here

Summary of input data

Name of input file:	
Specie and DB:	rno (rn4)
Number of allowed missmatches:	1
unique reads in input:	22086
number of reads in input (sum of all read counts)	1596647

Known MicroRNA

Library/ Parameters	mature	ambiguous mature	mature-star	ambiguous mature-star	unknown mature-star	ambiguos unknown mature-star	hairpin	ambiguous hairpin
total amount	128	29	25	2	14	0	59	11
fraction (amount) of known microRNAs	43.7% (293)	---	43.1% (58)	---	7.1% (197)	---	20.6% (287)	---
amount of unique reads	1219	176	111	9	54	0	216	18
fraction of unique reads	5.5%	0.797%	0.503%	0.041%	0.244%	0.000%	0.978%	0.081%
read count	1408255	44925	4115	88	1190	0	2823	111
fraction of read count	88.2%	2.814%	0.258%	0.006%	0.075%	0.000%	0.177%	0.007%
links to detail pages	details	details	details	details	details	no results	details	details

Alignment to transcriptome

Library/Parameters	Transcriptome	Rfam	RepBase	RepeatMasker (genomic)
amount of unique reads	1839	95	268	4906
fraction of unique reads	8.327%	0.430%	1.213%	22.213%
read count	47370	377	3950	19596
fraction of read count	2.967%	0.024%	0.247%	1.227%
links	details	details	details	details

Predicted candidate microRNAs

Number of predicted new microRNA: 91 (out of 93 predicted precursors)

112 from 22086 input reads (0.507 percent) sequence reads where found to be part of putative new microRNAs which corresponds to 422 expressed sequence reads out of 1596647 (0.026 percent)

See detailed analysis

Unmatched reads

Parameters	Filtered Reads	Unmapped Reads
amount of unique reads	0	8048
fraction of unique reads	0.000%	36.439%
read count	0	44942
fraction of read count	0.000%	2.815%
links	details	details

For questions or feedback please contact: Ana Maria Aransay or Michael Hackenberg

Figure 2. The summary page of miRanalyzer: five boxes are shown which correspond to summary & state of the process, analysis of known microRNA, matches against transcribed sequences, and detection of new microRNAs and summary of unmatched sequences.

Generating 'unknown mature-star' sequences

We generated the unknown star sequences by means of the mirBase precursor and mature sequences. First, we calculate the secondary structures for all hairpins using RNAfold (20) with parameters '-noLP'. Then, we detect the coordinates of the mature microRNAs within the pre-microRNA hairpin. By means of these coordinates, the information of the secondary structure and the characteristic '2-nt 3′ overhang' caused by Dicer, we extracted the corresponding sequence pairing with the mature microRNA.

Read alignment

Read sequences often contain adapter sequences (see standard protocol of small RNA sample preparation at http://www.illumina.com/) at its 3′ ends. Therefore, miRanalyzer has two alignment options depending on whether the reads have adapter sequences or not. In general, the tool generate a prefix tree of all input reads and subsequently walk in a single run over the genome to detect the reads. By default, miRanalyzer assumes the existence of adapter sequences and therefore, first detects matches of a subsequence of 16 bp starting at the 5′ end of the read. When miRanalyzer detects an initial match, it expands the subsequence as long as a perfect match is given. Finally, only matches of the longest subsequence are retained. Note that, in this approach the adapter sequences are detected implicitly (the sequence at the 3′ end of the read that does not match to the genome is defined as the adapter) and therefore, the adapter sequences need not to be known or supplied by the user.

Ontological analysis

We used a recently published tool, Annotation-Modules (14), to pre-calculate the significant annotations of all target gene lists for all microRNAs in the miRBase (12.0). Currently, the user can choose between two different target site prediction methods: miRBase target site predictions by miRanda software (21) and TargetScan (22).

Secondary structure prediction

For predicting the secondary structure and its minimum free energy (MFE) we utilized the Vienna RNA package (20).

The machine learning approach

To detect new microRNAs, we set up a machine learning approach based on the WEKA (23) implementation of the *random forest* learning scheme (16) with the number of trees set to 100. Note that, the random forest algorithm was applied by Jiang *et al.* (10) using basically the triplet structure features introduced by Xue *et al.* (9). However, the difference of our approach consists of using a negative set derived directly from the experimental data which (i) assures that the sequences are transcribed and (ii) allows the generation of new and previously unused features that seem to be more discriminative than the triplet structure features (see below).

Training and test sets

For the machine learning approach we created three data sets, one from each of the three species: human, *C. elegans* and rat. First, we extracted all pre-microRNA candidates from the experimental dataset that could be mapped to a known microRNA and labeled them as positive instances. Second, we selected an equal amount of pre-microRNA candidates from the same dataset by random selection with the known microRNAs removed and labeled them as negative. In total we obtained a dataset of 612 instances in human, 468 instances in worm and 376 instances in rat.

Features

We created a broad variety of features associated with nucleotide sequence, structure and energy. Table 3 lists all the features used in this work.

Table 3. Features calculated for the generation of the classifier

Feature name	Description of the feature
Read count	Number of reads mapping to the pre-microRNA
Length	The length of the longest hairpin structure
Stem length	The length of the longest hairpin structure stem
Mfe	The mean free energy of the hairpin
Loop length	The number of bases in the loop of the hairpin
Loop GC	The GC-content of the loop
GC	The GC-content of the small hairpin
Asymmetric bulges	The number of asymmetric bulges and mismatches regarding the stem
Symmetric bulges	The number of symmetric bulges and mismatches regarding the stem
Bulges	The number of bulges in the stem
Longest bulge	The number of non-pairing nucleotides of the longest bulge
Mismatches pre-microRNA	The number of single mismatches in the hairpin
Mismatches microRNAs	The number of single mismatches in the mature microRNA region of the hairpin
Stability	The smallest hairpin harbouring the read is extended 10 times 10bp at both ends. The stability is the frequency the original structure is found in the elongated structures
Alternating stability	Reports whether a structure disappears when extending the sequence, but reappears again.
Triplet-SVM features	All features that were proposed by Xue *et al.* (9)
Bindings	The number of bindings in the stem divided by the hairpin length

The selection of the features with highest prediction power was performed by means of calculating their information gain. Subsequently, we ranked the features according to their discrimination power. The top 10 features used for building the final classifier are: stability, mfe, bindings, stem length, read count, longest bulge, mismatches microRNA, mismatches pre-microRNA, alternating stability and the Triple-SVM feature 'A ...'. Supplementary Table S6 shows the 10 best features selected for each model used for data included in Table 2. It can be seen that nine features are always the same and just their ranking and the Triplet-SVM feature vary.

Pre-processing

In order to check the reads for putative new microRNAs we perform a pre-processing of the data which contains the following steps: (i) all reads which overlap in the genome are clustered together. (ii) Due to erroneous reads, dicer products (mature, mature-star and loop) could be grouped together such that they appear as non-microRNA products (for example producing a long cluster which overlaps the loop of the precursor). To avoid such a situation, we walk along the cluster sequences and test whether the start of the current read overlaps less than 3 nt with the end positions of previous reads. In that case the cluster is split at the current read start position. Clusters now contain a non-dicer product, the mature or the mature-star, but not more than one theoretical product. (iii) Clusters of more than 25 bp length are discarded. (iv) Since the microRNA can be located either on the 5′ arm or the 3′ arm of the hairpin, we extract the cluster sequence twice, with 60 bp upstream and 10 bp downstream flanking areas and vice versa. For both sequences the secondary structure is predicted via RNAfold, but only the energetically favourable is retained. (v) Non-hairpin structures are discarded. (vi) Structures where the cluster sequence is not fully included or spans the loop and a part of the stem cannot be dicer products are consequently discarded. Finally, since our analysis showed that virtually all known microRNAs show more than 14 bindings in the microRNA:microRNA-star duplex, we considered this as a mandatory requirement. Having applied the pre-processing step to the three experimental data sets, we receive 6967 candidate precursors for rat, 12 233 for worm and 43 905 for human.

Post-processing

After classification of the deep-sequencing data in form of the clusters created in the pre-processing step, clusters containing the mature and mature-star microRNA are merged such that one cluster represents one microRNA precursor.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dmitrij Frishman for careful reading of the manuscript and helpful comments, Philipp Pagel for helpful suggestions, Ewa Gubb for revising the English style and the Genome Analysis Platform staff at CIC bioGUNE for their technical support.

FUNDING

The Department of Industry, Tourism and Trade of the Government of the Autonomous Community of the Basque Country [Etortek Research Annual Programs]; the Innovation Technology Department of the Bizkaia County [2008–2009 institutional support to technological platforms]; Junta de Andalucia [P07FQM3163 to MH]. Funding for open access charge: [Etortek IE08-228].

Conflict of interest statement. None declared.

REFERENCES

1. Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
2. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
3. Ouellet,D.L., Perron,M.P., Gobeil,L.A., Plante,P. and Provost,P. (2006) MicroRNAs in gene regulation: when the smallest governs it all. *J. Biomed. Biotech.*, **2006**, 69616.
4. Bagasra,O. and Prilliman,K.R. (2004) RNA interference: the molecular immune system. *J. Mol. Histol.*, **35**, 545–553.
5. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
6. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
7. Friedländer,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotech.*, **26**, 407–415.
8. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
9. Xue,C., Li,F., He,T., Liu,G.P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
10. Jiang,P., Wu,H., Wang,W., Ma,W., Sun,X. and Lu,Z. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339–W344.
11. Nam,J.W., Kim,J., Kim,S.K. and Zhang,B.T. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.*, **34**, W455–W458.
12. Bentwich,I., Avniel,A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genetics*, **37**, 766–770.
13. Griffiths-Jones,S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol.*, **342**, 129–138.
14. Hackenberg,M. and Matthiesen,R. (2008) Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics*, **24**, 1386–1393.

15. Berezikov,E., van Tetering,G., Verheul,M., van de Belt,J., van Laake,L., Vos,J., Verloop,R., van de Wetering,M., Guryev,V., Takada,S. *et al.* (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.*, **16**, 1289–1298.
16. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 28.
17. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
18. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
19. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Res.*, **110**, 462–467.
20. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
21. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
22. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
23. Witten,I. and E,F. (2005) *Data Mining: practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco.