miRNAfe: a comprehensive tool for feature extraction in microRNA prediction

Cristian A. Yones[a,*], Georgina Stegmayer[a], Laura Kamenetzky[b], Diego H. Milone[a]

[a]*Research Center for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.*
[b]*Instituto de Investigaciones en Microbiología y Parasitología Médica (UBA), CONICET, Paraguay 2155, piso 13 (1121), Buenos Aires, Argentina.*

Abstract

miRNAfe is a comprehensive tool to extract features from RNA sequences. It is freely available as a web service, allowing a single access point to almost all state-of-the-art feature extraction methods used today in a variety of works from different authors. It has a very simple user interface, where the user only needs to load a file containing the input sequences and select the features to extract. As a result, the user obtains a text file with the features extracted, which can be used to analyze the sequences or as input to a miRNA prediction software.

The tool can calculate up to 80 features where many of them are multi-dimensional arrays. In order to simplify the web interface, the features have been divided into six pre-defined groups, each one providing information about: primary sequence, secondary structure, thermodynamic stability, statistical stability, conservation between genomes of different species and substrings analysis of the sequences. Additionally, pre-trained classifiers are provided for prediction in different species. All algorithms to extract the features have been validated, comparing the results with the ones obtained from software of the original authors.

The source code is freely available for academic use under GPL license at `http://sourceforge.net/projects/sourcesinc/files/mirnafe/0.90/`. A user-friendly access is provided as web interface at `http://fich.unl.edu.ar/sinc/web-demo/mirnafe/`. A more configurable web interface can be accessed at `http://fich.unl.edu.ar/sinc/web-demo/mirnafe-full/`.

Keywords: microRNA, Feature extraction, Web tool

---

*Corresponding author
  Email address:* `cyones@sinc.unl.edu.ar` (Cristian A. Yones)

1. Introduction

MicroRNAs (miRNA) are a group of short ($\sim$ 22 nucleotides) non-coding RNA which can play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Lamers et al., 2014). Precursors of miRNA
5 (pre-miRNA) are characterized by their hairpins structure. However, a large amount of similar sequences can be folded into this kind of structure in many genomes.

In order to predict miRNAs, a large number of tools have been developed in the last years (Kleftogiannis et al., 2013). The first step is to extract features
10 from sequences and then use classifiers to predict which sequences are likely to contain a miRNA. The feature extraction step is very important for the whole process, in order to achieve high rates of true positives predictions (Zhang et al., 2010). Numerous features can be extracted from the primary sequence and its corresponding secondary structure. A typical example of this kind of features
15 is the *triplets* representation (Xue et al., 2005), which considers the structural composition of three adjacent nucleotides and the middle base to build a vector with 32 elements. Other examples are the number of internal loops and their length (Yousef et al., 2006), the z-score of the minimum free energy (Hertel and Stadler, 2006) and the dinucleotide proportion (Rukshan and Vasile, 2009).
20 The amount of features that can be extracted is very large and there are many different tools that partially achieve this task. They are coded in different programming languages and have different access modes (web, command line, etc.). Besides, several tools are proprietary software and the source code is not even available[1]. These are important issues that hinder their use.
25 We have developed the miRNAfe tool that implements almost all existing state-of-the-art feature extraction processes used for miRNA prediction nowadays (Li et al., 2010). It can extract the features used by the most cited miRNA classifiers, such as Triplet-SVM (Xue et al., 2005), RNAmicro (Hertel and Stadler, 2006), BayesMiRNAfind (Yousef et al., 2006), MiRFinder (Huang
30 et al., 2007), MiPred (Jiang et al., 2007), miRRim (Goro et al., 2007), microPred (Rukshan and Vasile, 2009), miRanalyzer (Hackenberg et al., 2009), MiRenSVM (Jiandong et al., 2010) and miPredGA (Xuan et al., 2011). We have developed an easy to use web interface that allows a single and simplified access point to all the functions of the toolbox, and a set of pre-trained classifiers that can be
35 used to test the prediction power of the feature sets. We provide here a comprehensive open-source solution, with free access to all features for academic use.

2. Provided features

The tool implements up to 80 features, where many of them return arrays.
40 All of these features have been proposed in literature over the past 10 years.

---

[1]http://www.insybio.com/pages/ncrnaseq

The features are divided into six pre-defined groups according to the kind of information that must be extracted from the sequence. A brief explanation of each group is provided in the next lines. For a more detailed explanation of all the features provided and their sources, see the supplementary material.

2.1. Sequence

These are the simplest features and represent information from the primary sequence. MiRNAfe can extract a total of 5 features in this group: sequence length (ℓ), proportion of each base in the sequence, proportion of dinucleotides, content of guanine and cytosine and guanine-cytosine ratio. The last two features are defined as:

$$G + C_{content} = \frac{G + C}{G + C + A + U}, \tag{1}$$

$$GC_{ratio} = \frac{G}{C}, \tag{2}$$

where G, C, A and U represent the quantity of each base found in the sequence (Hertel and Stadler, 2006). All these features form a vector of 23 elements, composed by: the 4 base proportions, the 16 dinucleotide proportions, sequence length, $G + C_{content}$ and GC_{ratio}. Although these features are quite simple, they have shown a high discriminative power (Rukshan and Vasile, 2009), and thus are used in most of the state-of-the-art prediction software.

2.2. Secondary structure

These features represent information from the secondary structure and they are the most numerous group. The most used feature of this group is the triplets proportion (Xue et al., 2005). A triplet is an element formed with the structure state (paired or not paired) of three adjacent nucleotides and the base at the middle. An example of a triplet element is ".((A", where the parenthesis represents a paired nucleotide, a dot a not paired one, and the letter is the base of the nucleotide in the middle. As there are 2 possible states for a nucleotide and 4 different bases, 32 triplets can be formed (4×2^3). The number of occurrences of each triplet element in the sequence is counted and normalized to produce a 32-dimensional feature vector. A similar approach to the triplets was used by Huang et al. (2007), which proposed another representation for the secondary structure. First of all, five symbols are defined to indicate the status of each base pair in the stem: "=", ":", "−", "." and "∧". Each of them corresponds to the status of match, mismatch, deletion, insertion in the interior loop, and insertion in the bulged loop, respectively. Then, by taking two adjacent symbols, 14 possible combinations can be formed, each one having a special meaning. For example: "= −", "= .", and "=:" represent the boundary of the stem/loop, and "∶ ∧" represents that the loop is asymmetric. The frequency of each combination is used as a feature vector. This representation is also used to calculate four more features: $pMatch$, $pMismatch$, pDI and $pBulge$. These features are calculated over putative mature miRNA, selected as the 22 nucleotide region

where base-pairing is maximum. They represents the base pairing frequency, the non-pairing frequency, the deletion and insertion frequencies and the symmetry of the bulged loops, respectively.

Another kind of features is related to the stems, which are structural motifs containing more than three contiguous base pairs (Ng and Mishra, 2007). These features are the number of stems, the proportion of each possible base pair per stem, average base pair number per stem and length of the longest stem. The rest of features are the stem region (the stem part of the stem-loop) length, terminal loop length, bulges number, loops number, longest loop length, asymmetric and symmetric loops number, nucleotides in symmetric and asymmetric loops, longest symmetric region, average length of symmetric loops, average length of asymmetric loops, number of bulges and loops of length $1, 2, ..., 7$ and greater, base pair number, adjusted base pair propension, base pair proportion and $G+C_{content}$ in the terminal loop (Lopes et al., 2014). Finally, miRNAfe can calculate reads count from RNAseq data. This feature needs the user to provide an extra file with reads, which miRNAfe aligns with the analyzed sequences and counts the corresponding matches. For a full description of each feature see the supplementary material.

2.3. Thermodynamics stability

The features in this group are related to the thermodynamics stability of a sequence. The mostly used feature is the minimum free energy (MFE): the estimated energy that one sequence frees when folded into the most stable secondary structure (Zuker and Stiegler, 1981). The ensemble free energy (EFE) has a similar meaning and it is obtained with the algorithm from McCaskill (1990). Other features of this group are calculated as combinations of those values. For example, the MFE index 1 ($MFEI_1$) is the ratio between the minimum free energy and the $G + C_{content}$ defined in 1. Similarly, miRNAfe can calculate $MFE - EFE$ difference, adjusted MFE, $MFEI_2$, $MFEI_3$ and $MFEI_4$ (Rukshan and Vasile, 2009). There are also some features that use information theoretic approaches to estimate the confidence of the predicted secondary structure, such as the adjusted Shannon entropy of the pairing probabilities (Ng and Mishra, 2007), defined as

$$dQ = \frac{1}{\ell} \sum_{i<j} p_{ij} \log_2 p_{ij}, \tag{3}$$

where p_{ij} is the probability that the nucleotide i forms a pair with the nucleotide j and l is the sequence length. The base pair probabilities are calculated with the algorithm from McCaskill (1990). Another example is the adjusted base pair distance, defined as

$$dD = \frac{1}{\ell} \sum_{i<j} p_{ij}(1 - p_{ij}). \tag{4}$$

Additionally, in this group miRNAfe can calculate the ensemble frequency, set diversity, stem 3' and 5' potential, and loop potential (Terai et al., 2007).

There are 15 features in this group, which are described in more detail in the supplementary material.

2.4. Statistical stability

¹²⁰ It is well-known that precursors containing a miRNA are more stable than random sequences. The features in this group are calculated as the standard score of any feature related to stability. To calculate this score, a random population of sequences has to be generated swapping the bases of the analyzed sequence. This way, the artificially generated sequences conserve the nucleotide ¹²⁵ proportions or even the dinucleotide proportion if some swaps are restricted (the tool has an option to choose which swap method to use). For each generated sequence, the stability can be measured with the z-score (Bonnet et al., 2004), defined as

$$z = \frac{x - \mu}{\sigma}, \tag{5}$$

where x is the original value of the feature, μ is the mean and σ is the standard ¹³⁰ deviation of the randomly generated population of sequences. This score represents how many standard deviations a value is above the population mean. Thus, a negative z-score indicates a sequence that is statistically more stable that the population mean. Another statistic used to measure the stability of the sequence in comparison with random sequences is the p-value. It is calculated ¹³⁵ as the proportion of random sequences that are more stable that the analyzed sequence. Thus, a low p-value indicates that the analyzed sequence is one of the most stable of all sequences generated with that nucleotide/dinucleotide proportion. The stabilities measures that can be normalized with z-score are: MFE (named $zMFE$), EFE ($zEFE$), adjusted MFE (zG), Shannon's en-¹⁴⁰tropy (zQ), base pair propensy (zP) (Ng and Mishra, 2007) and base pair distance (zD) (Jiandong et al., 2010). The p-value can be used to normalize the MFE ($pMFE$) (Bonnet et al., 2004) and the EFE ($pEFE$) (Jiandong et al., 2010). Although z-score and p-value are alternative statistics for these features, they are often used together in prediction since they can take very different ¹⁴⁵values (Jiandong et al., 2010).

In summary, miRNAfe can calculate 8 features in this group. In the full version, the user can specify the shuffling method (preserving nucleotide or dinucleotide composition) and the number of random sequences generated. For the user-friendly web-interface, these parameters are set by default to 1000 ¹⁵⁰random sequences and preservation of dinucleotide composition.

2.5. Phylogenetic conservation

When a portion of the genome is conserved between related species, it is highly likely to have an important role in the genome. The features of this group measure the level of conservation between sequences of phylogenetically ¹⁵⁵related species. All the features are calculated over alignments of two or more sequences that the user must provide. Some features do not only take into account the conservation level, but also the thermodynamic stability. The features

5

in this group are: the mutation frequency (Huang et al., 2007), which is the proportion of bases that differ from one sequence to another and it is applicable
only to a pair of sequences; the column entropy of the 5' arm, 3' arm, loop region and minimum entropy, which is the Shannon entropy calculated over a region of 21 nucleotides (Hertel and Stadler, 2006); the number of differences in the secondary structure divided by the number of differences between sequences (Huang et al., 2007); the average MFE; the MFE difference between
two aligned sequences, divided by the number of differences between the sequences(Huang et al., 2007); average dG; average $MFEI_1$; free energy of the consensus secondary structure; conservation of the 3' arm and conservation of the 5' arm; and finally, the conservation score. This is the most complex feature to obtain (Goro et al., 2007), because is calculated using two Markov processes,
one that moves in the time dimension (over the branches of the evolution tree), and the other in space dimension (over the sequence). A total of 14 features can be extracted in this group, which are described in detail in the supplementary material.

2.6. 22-nt substring analysis

These features are calculated over all 22 nt substrings within a given sequence. They are based on the fact that if one sequence is a pre-miRNA, one of the analyzed substring has to be the mature miRNA and the features calculated must capture its particularities. As a result, an array with length $n = \ell - 22$ is obtained, where the i-th element represents the value of the calculated feature
over the substring that starts at the base i. MiRNAfe can extract the following 5 features in this group: the base-pairing probability in the substring (Lim et al., 2003), which is the sum of the base-pairing probability over the substring; the sum of not paired bases on the substring; the sum of the base-pairing probability on the secondary structure, without the probabilities of the nucleotides on
the substring; the bulge symmetry, as the difference between the amount of not paired bases on each arm of the substring; and the distance from the substring to the terminal loop.

3. Implementation

MiRNAfe is composed by a set of Matlab functions which prepare the input
sequences and implement the feature extraction processes. The source code is platform independent and provides functions that allow batch processing, saving of results and show reports with the extracted features. These functions can be installed in the user machine and used as any other Matlab toolbox. Thus the user is able to extract all the features present in miRNAfe and also make
predictions. The toolbox has a main function that takes as parameters the path of the input fasta file and a configuration file written in $yaml^2$, which is a human readable format that allows editing in a simple way on any text editor. This

---

[2]http://www.yaml.org/

file contains folding options, alignment parameters to make phylogenetic related tests, a list of features to extract, post-process options such as normalization or sequence filtering by minimum free energy. The toolbox is very versatile and can be extended easily, only saving the function in the source folder and adding one line to the configuration file. Also, miRNAfe when used as a toolbox can train a classifier and optimize its parameters. In this case, the user only needs to provide positive and negative examples of miRNA sequences of a target specie or family. In the configuration file, options related to the SVM training and the parameters optimization stage can also be specified. The toolbox uses external well-known software to implement some standard processes, like folding or aligning sequences. The Vienna RNA[3] package is used to fold the sequences and for alignments of sequences. Since the software RNAfold is used in almost all feature extraction process, miRNAfe shares its same restrictions about sequence lengths. Thus, a limit of 5000 nucleotides was imposed to avoid memory problems. For the phylogenetic related features, the software ClustalW[4] is used to align sequences, Bowtie[5] is used to align reads to sequences and PHAST[6] is used to calculate the conservation score.

4. Web interface

To provide a more user-friendly access, we have developed a simplified and easy to use web interface using the tool provided by Stegmayer et al. (2015). It can be accessed at `http://fich.unl.edu.ar/sinc/web-demo/mirnafe/`. As it can be seen in Figure 1, the user must load a fasta file with the sequences to be analyzed. After that, he/she can select which group of features wants to extract by checking the corresponding checkbox. Then, a pre-trained classifier on the sequences under analysis can be used for prediction. A support vector-machine (SVM) is provided for classification since it is the most frequently used method for pre-miRNA prediction (Kleftogiannis et al., 2013).Sequences from two genomes were used as training data, *Homo sapiens* and *Arabidopsis thaliana*. To create positive sets, all known pre-miRNAs from those species in miRBase release 21[7] (Kozomara and Griffiths-Jones, 2014) were used. Negative sets were built by extracting random sequences from the genomes and mRNAs of these species. The sequence length distribution in the negative dataset was the same as in the corresponding positive one. The extracted sequences were filtered to preserve only sequences with minimum free energy below -0.05 (normalized to the sequence length) and proportion of paired bases in the stem above 0.15, similarly to Gudyś et al. (2013). Since the features of the substring group are arrays of variable length, they cannot be used for prediction directly in a standard SVM. Similarly, since the features related to conservation are calculated

---

[3]`http://www.tbi.univie.ac.at/RNA/`
[4]`http://www.clustal.org/omega/`
[5]`http://bowtie-bio.sourceforge.net/index.shtml`
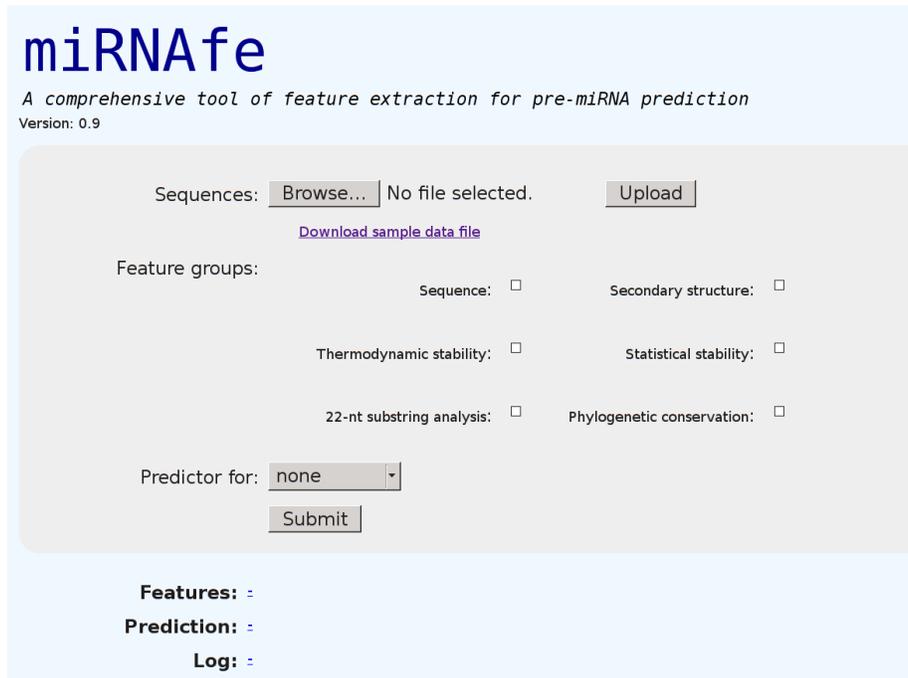[6]`http://compgen.bscb.cornell.edu/phast/index.php`
[7]`http://www.mirbase.org/`

Figure 1: Web interface of miRNAfe after analyzing some sample sequences

over several sequences altogether, they cannot be used directly for prediction with the pre-trained SVM.

Finally, when the feature extraction process is finished, the web interface provides links to download the output files in comma separated values (*csv*) format. This file format can be opened by any spreadsheet program and it is supported by almost all software used in bioinformatics. The features file contains, in each row, the name of the sequence analyzed, and in the first row the names of the corresponding features extracted. If the user has chosen to make a prediction, an extra file can be downloaded as well, which lists all input sequences ordered from most to less probable miRNA, with a flag indicating if it was classified as miRNA (positive flag) or not (negative flag). Finally, a log file is also provided, which describes all the process stages, warnings and errors, if any.

5. Validation of the feature extraction processes

In order to validate all feature extraction scripts, several sequences were analyzed with miRNAfe and with the software of the original authors, and the outputs have been compared. The sequences used in this validation step were the well-known pre-miRNAs: ppa-mir-101, hsa-mir-34a, hsa-mir-7-1, hsa-let-7a-1, hsa-let-7a-3 and hsa-let-7b3. These sequences were selected randomly with

8

²⁵⁵ the sole purpose of validating the feature extraction functions. The software used to validate the features extraction were MiRFinder (Huang et al., 2007), miPred (Jiang et al., 2007) and microPred (Rukshan and Vasile, 2009). MiRNAfe validates the extracted features by comparison with the reference software, when available. In most cases, however, the original scripts were not available ²⁶⁰ and reference results were obtained directly from data with already extracted features, and compared with the results of miRNAfe. In all tests performed, the results were always consistent to those of the original papers.

The validation process was automated with a script that is distributed together with the source code of the tool. This script prints on screen the results ²⁶⁵ of each test, the reference value and the software that was used to obtained it. This can also be used to make tests and re-validate the results of each function after making improvements or alternative implementations of the algorithms. For the statistical features, where results are not deterministic, a confidence interval for the expected values was set according to the variance in each case. ²⁷⁰ The validation error was calculated as $e = 100\|C - E\|_2/\|E\|_2$, where C is the result calculated by miRNAfe, E is the expected value (the one calculated with the software of the original author) and $\|\cdot\|_2$ is the norm 2. A list with the software used for comparisons and their corresponding references are provided in the supplementary material.

²⁷⁵ **6. Acknowledgments**

²⁸⁰ **References**

S. L. Lamers, G. B. Fogel, D. J. Nolan, M. S. McGrath, M. Salemi, HIV-associated neuropathogenesis: A systems biology perspective for modeling and therapy, Biosystems 119 (2014) 53–61.

D. Kleftogiannis, A. Korfiati, K. Theofilatos, S. Likothanassis, A. Tsakalidis, ²⁸⁵ S. Mavroudi, Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role, Journal of biomedical informatics 46 (3) (2013) 563–573.

X. Zhang, X. Song, H. Wang, Characteristic comparison between two types of miRNA precursors in metazoan species, Biosystems 100 (2) (2010) 144–149.

²⁹⁰ C. Xue, F. Li, T. He, G. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine, BMC Bioinformatics 6(1) (2005) 310.

M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. Showe, M. Showe, Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier, Bioinformatics 22 (11) (2006) 1325–1334.

J. Hertel, P. Stadler, Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data, Bioinformatics 22 (14) (2006) e197–e202.

B. Rukshan, P. Vasile, microPred: effective classification of pre-miRNAs for human miRNA gene prediction, Bioinformatics 25(8) (2009) 989–995.

L. Li, J. Xu, D. Yang, X. Tan, H. Wang, Computational approaches for microRNA studies: a review, Mamm Genome 21(1) (2010) 1–12.

T. Huang, B. Fan, M. Rothschild, Z. Hu, K. Li, S. Zhao, MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans, BMC Bioinformatics 8(1) (2007) 341.

P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, Z. Lu, MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features, Nucleic acids research 35 (suppl 2) (2007) W339–W344.

T. Goro, K. Takashi, A. Kiyoshi, K. Taishin, miRRim: A novel system to find conserved miRNAs with high sensitivity and specificity, RNA 13 (12) (2007) 2081–2090.

M. Hackenberg, M. Sturm, D. Langenberger, miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Research 37 (2009) 68–76.

D. Jiandong, Z. Shuigeng, G. Jihong, MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features, BMC Bioinformatics 11 (11) (2010) 11.

P. Xuan, M. Guo, J. Wang, C. Wang, X. Liu, Y. Liu, Genetic algorithm-based efficient feature selection for classification of pre-miRNAs, Genet. Mol. Res. 10 (2) (2011) 588–603.

K. Ng, S. Mishra, De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures, Bioinformatics 23(11) (2007) 1321–30.

I. Lopes, A. Schliep, A. C. d. L. de Carvalho, The discriminant power of RNA features for pre-miRNA recognition, BMC bioinformatics 15 (1) (2014) 124.

M. Zuker, P. Stiegler, Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information, Nucl Acid 9 (1981) 133–148.

J. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers 29 (1990) 1105–1119.

G. Terai, T. Komori, K. Asai, T. Kin, miRRim: a novel system to find conserved
miRNAs with high sensitivity and specificity, Rna 13 (12) (2007) 2081–2090.

E. Bonnet, J. Wuyts, P. Rouz, Y. Van de Peer, Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than
random sequences, Bioinformatics 20 (17) (2004) 2911–2917.

L. Lim, N. Lau, E. Weinstein, A. Abdelhakim, S. Yekta, M. Rhoades, C. Burge,
D. Bartel, The microRNAs of Caenorhabditis elegans, Genes & development
17 (8) (2003) 991–1008.

G. Stegmayer, M. Pividori, D. H. Milone, A very simple and fast way to access
and validate algorithms in reproducible research., Briefings in Bioinformatics.
Advance Access published July 28, 2015, doi: 10.1093/bib/bbv054.

A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence microRNAs using deep sequencing data, Nucleic Acids Research 42 (2014) D68–D73.

A. Gudyś, M. W. Szcześniak, M. Sikora, I. Makałowska, HuntMi: an efficient
and taxon-specific approach in pre-miRNA identification, BMC bioinformatics 14 (1) (2013) 83.