miRNAfe detailed feature list

Cristian A. Yones, Georgina Stegmayer, Laura Kamenetzky, and Diego H. Milone

1 Sequence

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Length of sequence	l		[3]$^{(ap)\,1}$	`sequence_length`	1
2. Nucleotide proportion	$A\%, C\%, G\%, U\%$	Ratio of each base in the sequence	[11]$^{(a)}$	`nt_proportion`	4
3. Dinucleotide ratio	$AA\%$, $AU\%$, ..., $GC\%$, $GG\%$	Ratio of dinucleotide elements of each kind.	[10]$^{(a)}$, [14]$^{(a)}$	`dinucleotide_proportion`	16
4. G+C content	-	Aggregated proportion of guanine and cytosine on the sequence $G+C_{content} = \frac{G+C}{G+C+A+U}$	[4]$^{(a)}$, [10]$^{(a)}$, [3]$^{(ap)}$, [14]$^{(a)}$	`gc_content`	1
5. G/C ratio	-	Ratio of guanine over cytosine $G/C_{ratio} = \frac{G}{C}$	[6]$^{(a)}$	`gc_ratio`	1

[1] the features of this reference were used in: (a) animals, (p) plants and/or (v) viruses.

2 Secondary structure

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Triplets	-	Vector of 32 elements with the triplets frequency. A triplet is an element formed with the structure composition (paired or not paired) of three adjacent nucleotides and the base of the middle. An example of these elements is ".((A", where the parenthesis represent a paired nucleotide, a dot a not paired one and the letter is the base of the middle nucleotide	[15](apv), [9](ap), [3](ap), [6](a)	`triplets`	32
2. Huang elements proportion	-	This feature uses Huang's notation. It is a vector with 10 elements where each one is the proportion of a Huang element ("=-", "==", "=:", "_", "-=", "^^", "^=", "::", ":^" and ":=").	[5](a)	`huang_elements_proportion`	10
3. Huang's $pMatch$ **ratio**	$pMatch$	This feature use Huang's notation. $pMatch$ indicates the base pairing and is calculated over putative mature miRNA, selected as the 22 nucleotide region where it is maximum.	[5](a)	`huang_ratios`	1
4. Huang's $pMismatch$ **ratio**	$pMismatch$	This feature use Huang's notation and is calculated over putative mature miRNA, selected as the 22 nucleotide region where $pMatch$ is maximum. $pMismatch$ represents the frequency of non-pairing base pairs (indicated by the size of the interior loops).	[5](a)	`huang_ratios`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
5. Huang's pDI ratio	pDI	This feature use Huang's notation and is calculated over putative mature miRNA, selected as the 22 nucleotide region where $pMatch$ is maximum. pDI represents the deletion and insertion frequencies.	[5](a)	`huang_ratios`	1
6. Huang's $pBulge$ ratio	$pBulge$	This feature use Huang's notation and is calculated over putative mature miRNA, selected as the 22 nucleotide region where $pMatch$ is maximum. $pBulge$ indicates the symmetry of the bulged loops.	[5](a)	`huang_ratios`	1
7. Steam number	l_s	Number of stems in the secondary structure.	[4](a), [13](a), [3](ap)	`stem_number`	1
8. $A-U$ base pair proportion per stem	$A-U/N_{stems}$	Number of adenine-uracil base pair divided by the number of stems.	[11](a), [10](a), [6](a), [14](a)	`bp_proportion_stem`	1
9. $G-C$ base pair proportion per stem	$G-C/N_{stems}$	Number of guanine-cytosine base pair divided by the number of stems.	[11](a), [10](a), [6](a), [14](a)	`bp_proportion_stem`	1
10. $G-U$ base pair proportion per stem	$G-U/N_{stems}$	Number of guanine-uracil base pair divided by the number of stems.	[11](a), [10](a), [6](a), [14](a)	`bp_proportion_stem`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
11. Average base pair per stem	Avg_BP_Stem	Average of nucleotides per stem.	[11]$^{(a)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`avg_bp_stem`	1
12. Length of the longest stem	-	Longest region where the pairing is perfect.	[11]$^{(a)}$	`longest_stem_length`	1
13. Steam region length	l_s	Number of nucleotides in the stem region of the secondary structure.	[4]$^{(a)}$, [13], [3]$^{(ap)}$	`stem_length`	1
14. Terminal loop length	l_h	Amount of nucleotides not paired in the terminal loop of the secondary structure $l_h = l - l_s.$	[4]$^{(a)}$, [13]$^{(a)}$, [3]$^{(ap)}$	`terminal_loop_length`	1
15. Bulges number	N_b		[16]$^{(apv)}$	`bulge_number`	1
16. Loop number	N_l	Total number of loops, including the terminal loop.	[16]$^{(apv)}$, [3]$^{(ap)}$	`loops_number`	1
17. Longest loop length	l_{ll}		[3]$^{(ap)}$	`longest_loop_length`	1
18. Asymmetric loops number	N_{al}		[16]$^{(apv)}$	`aloops_number`	1
19. Symmetric loops number	N_{sl}		[3]$^{(ap)}$	`sloops_number`	1
20. Nucleotides in symmetric loops	N_{nsl}		[11]$^{(a)}$, [3]$^{(ap)}$	`nt_sloops`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
21. Nucleotides in asymmetric loops	N_{nsl}		[11]$^{(a)}$	`nt_aloops`	1
22. Longest symmetric region	-	Length and distance to terminal loop of the symmetric region without asymmetric loops or bulges. The symmetric loops are allowed.	[11]$^{(a)}$	`longest_simmetric_region`	1
23. Average length of symmetric loops	-		[11]$^{(a)}$	`avg_length_sloops`	1
24. Average length of asymmetric loops	-		[11]$^{(a)}$	`avg_length_aloops`	1
25. Number of bulges of length 1 to 7 and >7	-	Vector with the number of bulges of length 1, 2, ..., 7 and greater than 7.	[16]$^{(apv)}$	`nbulge_length`	8
26. Number of loops of length 1 to 7 and >7	-	Vector with the number of loops of length 1, 2, ..., 7 and greater than 7.	[16]$^{(apv)}$	`nloops_length`	8
27. Base pair number	nP	Number of base pair, i.e number of paired nucleotides divided by 2	[16]$^{(apv)}$	`bp_number`	1
28. Adjusted base pair propension	dP	Number of base pair divided by the nucleotide number.	[8], [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`dP`	1
29. $A-U$ pair proportion	$A-U\%$	Proportion of adenine-uracil over the total number of base pairs.	[11]$^{(a)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`bp_proportion`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
30. $G-C$ **pair proportion**	$G-C\%$	Proportion of guanine-cytosine over the total number of base pairs.	[11]$^{(a)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`bp_proportion`	1
31. $G-U$ **pair proportion**	$G-U\%$	Proportion of guanine-uracil over the total number of base pairs.	[11]$^{(a)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`bp_proportion`	1
32. $G+C$ **content in the terminal loop**	-	Aggregated proportion of guanine and cytosine on the terminal loop.	[3]$^{(ap)}$	`gc_content_loop`	1
33. Reads count	-	The number of reads that match with the stem region of the analyzed sequence.	[3]$^{(ap)}$	`read_count`	1

3 Thermodynamic stability

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Minimum free energy	MFE	Minimum free energy obtained with the algorithm from Zuker, M. y P. Stiegler, 1981.	[11]$^{(a)}$, [9]$^{(ap)}$, [5]$^{(a)}$, [3]$^{(ap)}$	`mfe`	1
2. Ensemble free energy	EFE	Ensemble free energy obtained with the algorithm from McCaskill, 1990.	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`efe`	1
3. MFE index 1	$MFEI_1$	Ratio between the minimum free energy and the G+C content.	[17], [8]$^{(ap)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`mfei1`	1
4. Difference of MFE and EFE	$Diff$	Difference between these two values, divided by the sequence length, $$Diff = \frac{MFE - EFE}{l}.$$	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`mfe_efe_difference`	1
5. adjusted MFE	dG	Minimum free energy divided by the sequence length.	[2]$^{(a)}$, [17], [8]$^{(ap)}$, [6]$^{(a)}$, [14]$^{(a)}$	`dG`	1
6. MFE index 2	$MFEI_2$	Ratio between the dG and the number of stems.	[8]$^{(ap)}$, [10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`mfei2`	1
7. MFE index 3	$MFEI_3$	Ratio between the dG and number of loops.	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`mfei3`	1
8. MFE index 4	$MFEI_4$	Ratio between the dG and the G+C content.	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`mfei4`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
9. Adjusted Shannon's entropy	dQ	Characterize the probability of base pairing in a secondary structure as a chaotic dynamic system $$dQ = \frac{1}{l}\sum_{i<j} p_{ij}\log_2 p_{ij},$$ where p_{ij} is the probability of pairing of nucleotides i and j. This value is calculated with the algorithm from McCaskill, 1990. Low values of dQ correspond to distributions dominated by a few bases likely to be matched. These bases are better predicted than those that have multiple alternative states.	[8]$^{(ap)}$, [10]$^{(a)}$, [14]$^{(a)}$	`dQ`	1
10. Adjusted base pair distance	dD	It is the base pair distance for all pairs of structures inferred from the sequence $$dD = \frac{1}{l}\sum_{i<j} p_{ij}(1-p_{ij}),$$	[8]$^{(ap)}$, [10]$^{(a)}$, [14]$^{(a)}$	`dD`	1
11. Ensemble frequency in the set	Freq	Obtained with the algorithm from McCaskill, 1990.	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`ensemble_frequency`	1
12. Set diversity	Diversity	Obtained with the algorithm from McCaskill, 1990.	[10]$^{(a)}$, [6]$^{(a)}$, [14]$^{(a)}$	`diversity`	1
13. Stem 5' potential	P^L	It is the maximum probability of pairing a nucleotide with other that is on the 5' direction. $$Pl_i = \max_{j<i} p_{ij},$$ where p_{ij} is the same defined for dQ.	[2]$^{(a)}$	`stem5_potential`	Variable

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
14. Stem 3' potential	P^R	It is the maximum probability of pairing a nucleotide with other than the corresponding in the 3' direction. $$Pl_i = \max_{j>i} p_{ij},$$ where p_{ij} is the same defined for dQ.	[2](a)	`stem5_potential`	Variable
15. Loop potential	V'	It is a vector where each element measures how likely a nucleotide can be part of the terminal loop $$V'_i = \sum_j \omega_{i-j} \left[\sum_k p_{j-k,j+k} + p_{j-k+1,j+k} \right]$$ where p_{ij} is the same defined for dQ and ω is a smoothing window.	[2](a)	`loop_potential`	Variable

4 Statistical stability

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Standard score of the MFE	$zMFE$	Minimum free energy normalized with z-score.	$[4]^{(a)}$, $[2]^{(a)}$	`zMFE`	1
2. Standard score of the EFE	$zEFE$	Ensemble free energy normalized with z-score.	$[6]^{(a)}$	`zEFE`	1
3. Standard score of the dG	zG	Adjusted minimum free energy normalized with z-score.	$[8]^{(ap)}$, $[10]^{(a)}$, $[14]^{(a)}$	`zG`	1
4. Standard score of the Shannon's entropy	zQ	Adjusted Shannon's entropy normalized with z-score.	$[8]^{(ap)}$, $[10]^{(a)}$, $[14]^{(a)}$	`zQ`	1
5. Standard score of the base pair propention	zP	Base pair propention adjusted and normalized using z-score.	$[8]^{(ap)}$, $[10]^{(a)}$, $[14]^{(a)}$	`zP`	1
6. Standard score of the base pair distance	zD	Adjusted base pair distance normalized using z-score.	$[6]^{(a)}$	`zD`	1
7. Monte Carlo and randomization test over MFE	$pMFE$	p-value of the ensemble free energy.	$[1]^{(ap)}$	`pMFE`	1
8. Monte Carlo and randomization test over EFE	$pEFE$	p-value of the minimum free energy.	$[6]^{(a)}$	`pEFE`	1

5 Phylogenetic conservation

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Mutation frequency	–	Number of mutation (differences) between two sequences of RNA. Only applicable to alignments of two sequences.	[5](a)	`mutation_frequency`	1
2. Column entropy of the 5' arm	$S5'$	Shannon's entropy of the 5' arm	[4](a)	`column_entropy`	1
3. Column entropy of the 3' arm	$S3'$	Shannon's entropy of the 3' arm	[4](a)	`column_entropy`	1
4. Column entropy of the loop region	$S0$	Shannon's entropy of the terminal loop	[4](a)	`column_entropy`	1
5. Minimum entropy	S_{min}	Minimum entropy calculated over a region of 21 nucleotides.	[4](a)	`column_entropy`	1
6. Secondary structure differences	$Vstrc$	Difference between the secondary structures of two sequences caused by mutations divided by the number of differences between sequences	[5](a)	`se_difference`	1
7. Average minimum free energy	$\bar{E}$	Mean of the minimum free energies of the sequences that are part of the alignment.	[4](a)	`mean_mfe`	1
8. MFE difference	$VMFE$	Difference between the minimum free energy of two aligned sequences divided by the number of differences between the sequences.	[5](a)	`mfe_difference`	1

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
9. Average of dG	$\bar{\epsilon}$	Mean of the adjusted minimum free energies of aligned sequences.	[4](a)	`mean_dG`	1
10. Average of $MFEI_1$	$\bar{\eta}$	Mean of the MFE_1 of the aligned sequences.	[4](a)	`mean_mfei1`	1
11. Free energy of the consensus secondary structure	E_{cons}		[4](a), [7](a)	`mfe_consensus`	1
12. Conservation of the 3' arm	–	Number of bases conserved in two or more sequences in the 3' arm, without the 10 first bases of the substring.	[4](a), [7](a)	`conservation_3`	1
13. Conservation of the 5' arm	–	Number of bases conserved in two or more sequences in the 5' arm, without the 10 first bases of the substring.	[7](a)	`conservation_5`	1
14. Conservation score	CS	Conservation score of the alignment of sequences. Internally uses the software PhyloFit[2]. This score 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).	[2](a) [12]	`conservation_score`	1

[2]http://compgen.bscb.cornell.edu/phast/index.php

6 22-nt substring analysis

Feature name	Abbreviation	Brief description	Reference	miRNAfe function name	Vector length
1. Base pair probability	–	Sum of base-pairing probability over the substring.	[7]$^{(a)}$	`ss_base_pair`	Variable
2. Not paired bases	–	Sum of not paired bases on the substring.	[3]$^{(ap)}$	`ss_not_paired`	Variable
3. Extension base pair probability	–	Sum of base-pairing probability on the secondary structure, without probabilities of the nucleotides on the substring.	[7]$^{(a)}$	`ss_extension_base_pair`	Variable
4. Bulge symmetry	–	The difference between the amount of not paired bases on each arm of the substring.	[7]$^{(a)}$	`ss_bulge_simetry`	Variable
5. Terminal loop distance	–	Distance from the substring to the terminal loop.	[7]$^{(a)}$, [13]$^{(a)}$	`ss_loop_distance`	Variable

miRNAfe validation of feature extraction processes

Cristian A. Yones, Georgina Stegmayer, Laura Kamenetzky, and Diego H. Milone

In the next table, the software used for comparisons and their corresponding references are presented.

Feature	Software used
Triplets	MiRFinder [5]
Huang ratios	MiRFinder [5]
Huang elements proportion	MiRFinder [5]
$G + C_{content}$	microPred [10]
Dinucleotide proportion	microPred [10]
$MFEI_1$	genRNAStats and RNAspectral of miPred [6]
$MFEI_2$	genRNAStats and RNAspectral of miPred [6]
$MFEI_3$	microPred [10]
$MFEI_4$	genRNAStats and RNAspectral of miPred [6]
MFE difference	MiRFinder [5]
Secondary structure difference	MiRFinder [5]
Mutation frequency	MiRFinder [5]
$zMFE$	genRandomRNA of miPred [6]
$zEFE$	genRandomRNA of miPred [6]
zQ	genRandomRNA of miPred [6]
zP	genRandomRNA of miPred [6]
zG	genRandomRNA of miPred [6]
zD	genRandomRNA of miPred [6]
$pEFE$	genRandomRNA of miPred [6]
$pMFE$	genRandomRNA of miPred [6]

References

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

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

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

[4] Hertel, J and PF Stadler: *Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data.* Bioinformatics, 22 (14):e197–e202, 2006.

[5] Huang, TH, B Fan, M Rothschild, ZL Hu, K Li, and SH Zhao: *MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans.* BMC Bioinformatics, 8(1):341, 2007.

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

[7] Lim, LP, NC Lau, EG Weinstein, A Abdelhakim, S Yekta, MW Rhoades, CB Burge, and DP Bartel: *The microRNAs of caenorhabditis elegans.* Genes & development, 17(8):991–1008, 2003.

[8] Ng, KLS and SK Mishra: *De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures.* Bioinformatics, 23(11):1321–30, 2007.

[9] Peng, J, W Haonan, W Wenkai, M Wei, S Xiao, and L Zuhong: *MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.* Nucleic Acids Research, 35:339–44, 2007.

[10] Rukshan, B and P Vasile: *micropred: effective classification of pre-miRNAs for human miRNA gene prediction.* Bioinformatics, 25(8):989–995, 2009.

[11] Sewer, A, N Paul, P Landgraf, A Aravin, S Pfeffer, MJ Brownstein, T Tuschl, E van Nimwegen, and M Zavolan: *Identification of clustered microRNAs using an ab initio prediction method.* BMC Bioinformatics, 6:267, 2005.

[12] Siepel, A and D Haussler: *Phylogenetic hidden Markov models.* In *In statistical methods in molecular evolution*, pages 325–351. Springer, 2005.

[13] Snorre, AH, S Ola, and S Pal: *Reliable prediction of drosha processing sites improves microRNA gene prediction.* Bioinformatics, 23(2):142–149, 2007.

[14] Xuan, P, MZ Guo, J Wang, CY Wang, XY Liu, and Y Liu: *Genetic algorithm-based efficient feature selection for classification of pre-miRNAs.* Genet. Mol. Res., 10 (2):588–603, 2011.

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

[16] Yousef, M, M Nebozhyn, H Shatkay, S Kanterakis, LC Showe, and MK Showe: *Combining multi-species genomic data for microRNA identification using a naive bayes classifier.* Bioinformatics, 22 (11):1325–1334, 2006.

[17] Zhang, BH, XP Pan, SB Cox, GP Cobb, and TA Anderson: *Evidence that miRNAs are different from other RNAs.* Cell. Mol. Life Sci., 63(2):46–254, 2006.