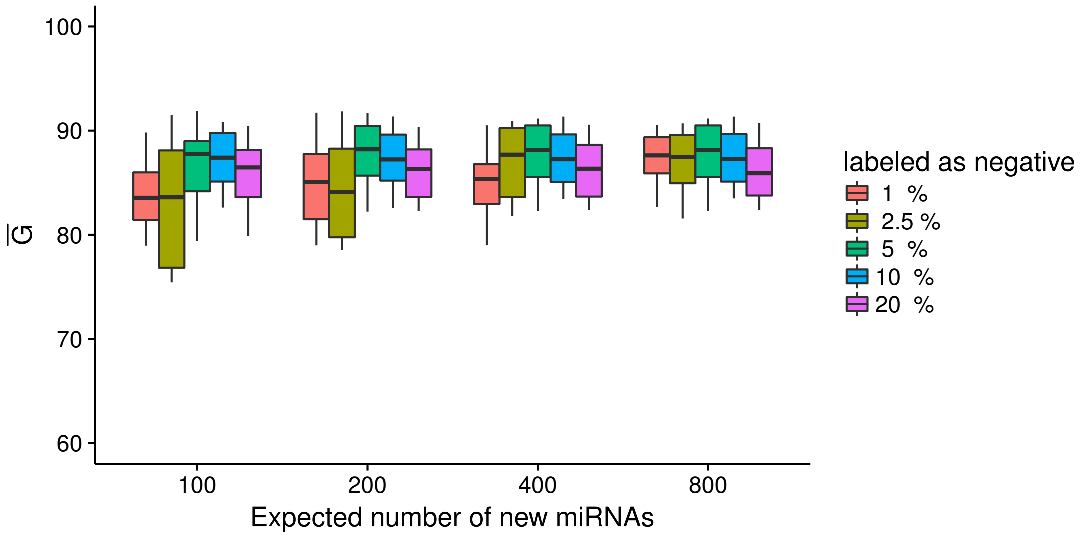
# Supplementary Material
Genome-wide pre-miRNA discovery from few labeled examples

C. Yones, G. Stegmayer and D. H. Milone

Research Institute for Signals, Systems and Computational Intelligence, sinc($i$), FICH-UNL, CONICET, Santa Fe, Argentina.

## Supplementary Figure S1: parameters sensitivity test



In this experiment, 20 combinations of the expected number of pre-miRNA to be find and the percentage of hairpins that will be labeled as negative examples to initiate the prediction algorithm have been tested on a cross-validation scheme, with the genome-wide dataset of *C. elegans*. The real number of true pre-miRNAs is 249, but as can be seen, for any estimation between 100 and 800 miRNAss maintains a good and stable performance. The same happens with the percentage of negative examples automatically labeled.

## Supplementary Section S2: avoiding misclassification of labeled examples

To avoid the misclassification of positive examples, the constant $c$ must be large enough to ensure that any misclassification of positive examples would yield a greater penalization than the regularizing term of the objective function. This value can be estimated from the equations of the method. Given the definition

$$a_{ij} = \begin{cases} \frac{\mu}{\mu+||\mathbf{x}_i-\mathbf{x}_j||^2} & \text{if } \mathbf{x}_j \in \mathcal{K}(\mathbf{x}_i) \text{ and } \ell_i\ell_j \geq 0 \\ 0 & \text{in other cases,} \end{cases} \tag{S.1}$$

since the norm cannot be negative, then $a_{ij} \leq 1$. Therefore, $d_{ii} = \sum_{k=0}^{n} a_{ik} \leq k$. From (2) in the manuscript,

$$\begin{aligned} \mathbf{z}^T L\mathbf{z} &= \mathbf{z}^T I\mathbf{z} - \mathbf{z}^T D^{-1/2^T} A D^{-1/2}\mathbf{z} \\ &= \sum_i^n z_i^2 - \sum_i^n \sum_j^n \frac{z_i}{\sqrt{d_{ii}}} \frac{z_j}{\sqrt{d_{jj}}} a_{ij} \\ &= n - \sum_i^n \sum_j^n z_i z_j \frac{a_{ij}}{\sqrt{d_{ii}}\sqrt{d_{jj}}} \\ &\leq n - \sum_i^n \sum_j^n z_i z_j \frac{1}{k} \\ &= n - \sum_i^n \frac{z_i}{k} \sum_j^n z_j = n - \sum_i^n \frac{z_i}{k} 0 = n \end{aligned} \tag{S.2}$$

Then, the misclassification of a positive example must have a penalization greater than $n$. A positive example $\mathbf{x}_i$ is misclassified when $z_i \leq 0$, which leads to

$$c(z_i - \ell_i)C_{ii}(z_i - \ell_i) \geq c(0 - \ell_i)C_{ii}(0 - \ell_i) = c\gamma_+ C_{ii}\gamma_+. \tag{S.3}$$

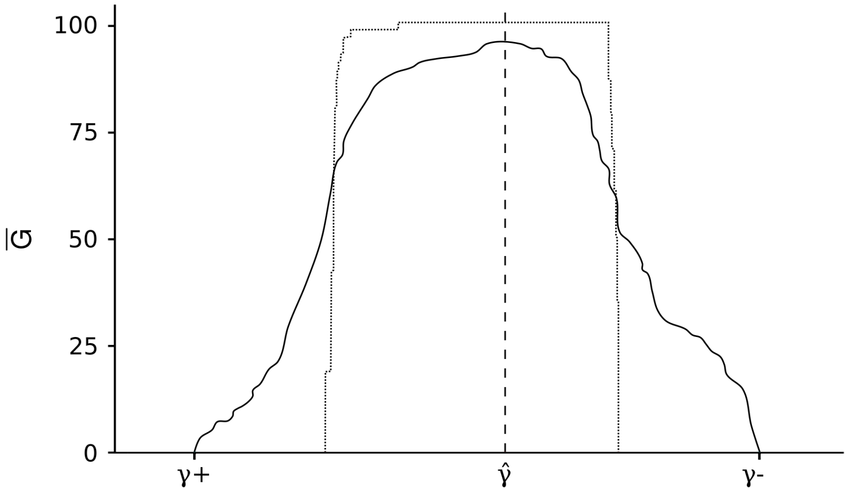
Then, to avoid the misclassification

$$\begin{aligned} c\gamma_+ C_{ii}\gamma_+ &> n \\ c\sqrt{\frac{n_-}{n_+}} C_{ii} \sqrt{\frac{n_-}{n_+}} &> n \\ c\frac{n_-}{n_+} C_{ii} &> n \\ cC_{ii} &> \frac{n_+ n}{n_-} \end{aligned} \tag{S.4}$$

Therefore, any combination of $c$ and $C_{ii}$ that fulfills the inequality S.4 would avoid the misclassification of positive examples. One option is to leave $C_{ii} = 1$ as default, and set

$c > \frac{n_+ n}{n_-}$. This parameter has the disadvantage that it also modifies the penalization for the negative examples. Another better solution is to leave $c = 1$ as default, and set $C_{ii} > \frac{n_+ n}{n_-}$ for the positive examples. With this setting only the positive examples are protected from misclassification.

## Supplementary Figure S3: thresholding the prediction scores



Comparison of the estimated (dotted line) and the real (solid line) geometric mean ($\bar{G}$) of sensitivity and specificity in an example dataset. Between two consecutive labeled samples in $\mathbf{z}^\star$ increasingly sorted by $\bar{G}$, there could be many unlabeled sequences. Hence, the estimated performance measure remains constant in those regions. When the number of labeled samples is low, this regions can be quite wide, therefore the final threshold ($\hat{\gamma}$) is set as the midpoint between the highest and the lowest scores in $\mathbf{z}^\star$ which maximizes the performance measure.

## Supplementary Section S4: feature sets

# Feature set 1 (FS1)

- $tri_A$, $tri_U$, $tri_G$, and $tri_C$: frequencies of secondary structure triplets composed of three adjacent nucleotides and the middle nucleotide: "A(((", "U(((", "G(((", and "C(((".

- $orf$: the maximal length of the amino acid string without stop codons found in three reading frames.

- $loops$: the cumulative size of internal loops found in the secondary structure.

- $dm$: a percentage of low complexity regions detected in the sequence using Dustmasker [1]

- $\%C + G$: aggregated proportion of cytosine and guanine on the sequence.

- $dG$: Minimum free energy divided by the sequence length.

- $dQ$: is calculated as
$$\frac{1}{l} \sum_{i<j} p_{ij} \log_2 p_{ij},$$
where $p_{ij}$ is the probability of pairing of nucleotides $i$ and $j$. This value is calculated with the software RNAfold with -p option. 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.

- $dF$: topological descriptor. For a further description see Gan *et. al.* (1987) [2].

- $MFEI_1$: ratio between the minimum free energy and the $\%C + G$.

- $MFEI_2$: is calculated as $dG/N_s$, where $N_s$ is the number of stems.

- $MFEI_3$: is calculated as $dG/N_l$, where $N_l$ is the number of loops in the secondary structure.

- $MFEI_4$: is calculated as $MFE/N_b$, where $N_b$ is the total number of base pairs in the secondary structure.

[1] Morgulis, A., Gertz, E. M., Schffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology*, **13**(5), 1028-1040.

[2] Gan, H. H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., ... & Schlick, T. (1987). RAG: RNA-As-Graphs databaseconcepts, analysis, and features. *Nutrition and Health*, **5**(1-2), 1285-1291.

- $zD$: base pair distance for all pair of structures, calculated as

$$dD = \frac{1}{l} \sum_{i<j} p_{ij}(1 - p_{ij}),$$

normalized with z-score.

- $Diversity$: the structural diversity calculated with RNAfold (-p option).
- $NEFE$: Normalized Ensemble Free Energy calculated with RNAfold (-p option).
- $Diff$: is calculated as $|MFE - EFE|/L$.
- $dS$: Structure Entropy calculated using UNAfold.
- $dS/L$: normalized structure entropy.
- $|A - U|/L$, $|G - C|/L$, $|G - U|/L$: number of each possible base pair normalized by the sequence length.
- $Avg\_BP\_Stem$: Average of nucleotides per stem.
- $\%(A-U)/N_s$, $\%(G-C)/N_s$ and $\%(G-U)/N_s$: proportion of base pairs on stems.

## Feature set 2 (FS2)

- Nucleotide proportion: ratio of each base in the sequence
- Dinucleotide proportion: ratio of dinucleotide elements of each kind.
- $L$: sequence length.
- $N_s$: number of stems.
- $\%C + G$: aggregated proportion of cytosine and guanine.
- $G/C_{ratio}$: ratio of guanine over cytosine.
- $Avg\_BP\_Stem$: average of nucleotides per stem.
- Longest stem length: longest region where the pairing is perfect.
- Terminal loop length: number of nucleotides in the stem region of the secondary structure.
- Number of base pair: number of paired nucleotides divided by 2
- $dP$: number of base pair divided by the nucleotide number.

- $|A-U|/L$, $|G-C|/L$, $|G-U|/L$: number of each possible base pair normalized by the sequence length.

- $\%(A-U)/N_s$, $\%(G-C)/N_s$ and $\%(G-U)/N_s$: proportion of base pairs on stems.

- 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.

- $MFE$: minimum free energy.

- $EFE$: ensemble free energy.

- $Freq$: the structural frequency calculated with RNAfold (-p option).

- $Diversity$: the structural diversity calculated with RNAfold (-p option).

- $Diff$: is calculated as $|MFE - EFE|/L$.

- $dG$: Minimum free energy divided by the sequence length.

- $dQ$: is calculated as

  $$\frac{1}{l}\sum_{i<j} p_{ij}\log_2 p_{ij},$$

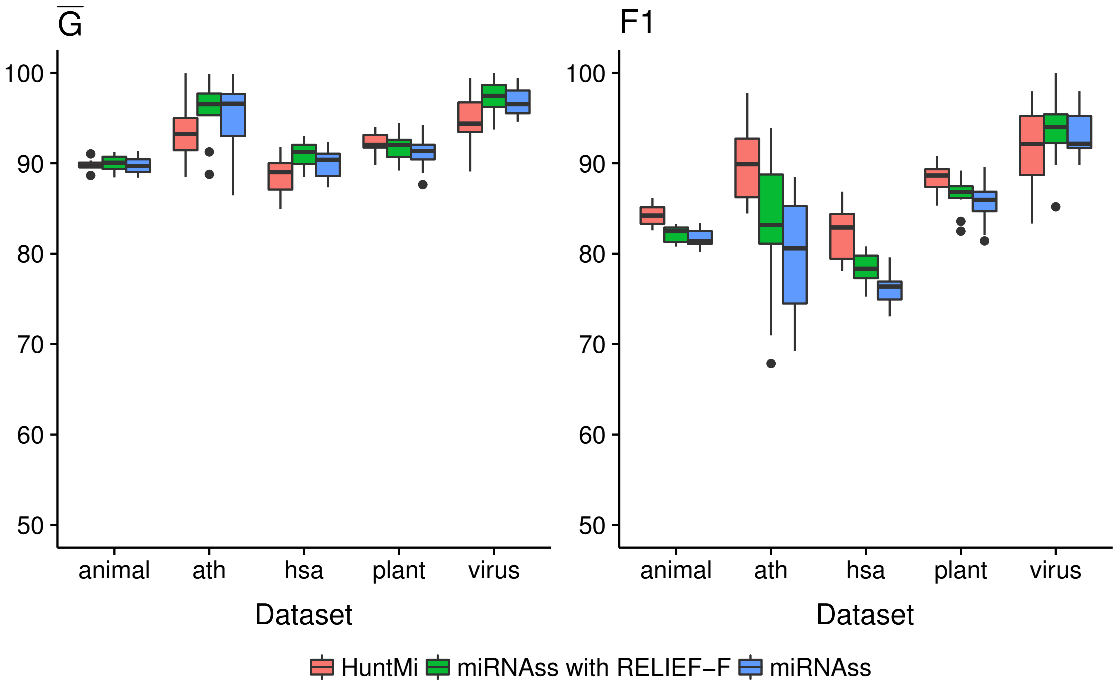
  where $p_{ij}$ is the probability of pairing of nucleotides $i$ and $j$. This value is calculated with the software RNAfold (-p option). Low values of $dQ$ correspond to distributions dominated by a few bases likely to be matched. These bases are better predicted than those that have multiple alternative states.

- $MFEI_1$: ratio between the minimum free energy and the $\%C + G$.

- $MFEI_2$: is calculated as $dG/N_s$, where $N_l$ is the number of stems.

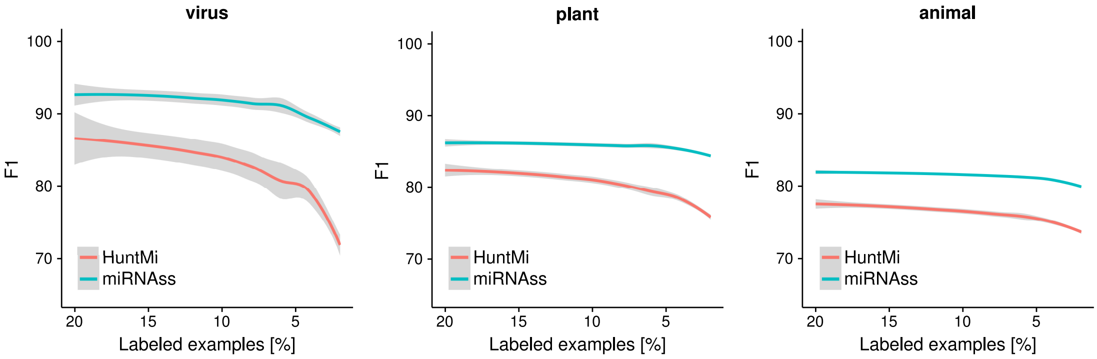
- $MFEI_4$: is calculated as $MFE/N_b$, where $N_b$ is the total number of base pairs in the secondary structure.

## Supplementary Figure S5: cross-validation in full labeled datasets



Box plot of $\bar{G}$ and $F_1$ obtained by miRNAss and HuntMi in five different datasets. The middle line in each box represents the median of each distribution. The upper whisker extends from the hinge to the highest value, that is, within 1.5 times the interquartile range ($IQR$) of the hinge. The lower whisker extends from the hinge to the lowest value, within 1.5 times $IQR$ of the hinge. Data beyond the end of the whiskers are outliers and are plotted as points.

## Supplementary Figure S6: few labeled examples



Curves of $F_1$ obtained by decreasing the percentage of labeled sequences. The shaded regions are confidence intervals of the estimation with local regression (LOESS) at $p < 0.05$. In the Virus dataset, while the $F_1$ achieved by HuntMi falls as the percentage of labeled examples decreases, miRNAss maintains a higher and almost constant $F_1$, independently of the percentage of labeled sequences. In the plant dataset, again, miRNAss achieves a higher $F_1$ for all percentages; and moreover, this difference increases as the number of labeled examples decreases. In animal dataset, miRNAss maintains an almost constant $F_1$ when the percentage of labeled sequences is greater than 5%. HuntMi achieves a lower $F_1$ for all percentages.