miRNAss: a semi-supervised approach for microRNA prediction

Cristian Yones, Georgina Stegmayer, Diego Milone
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, (3000) Santa Fe, Argentina.

Background

MicroRNA (miRNAs) play essential roles in post-transcriptional gene regulation in animals and plants. Precursors of miRNA (pre-miRNA) are characterized by their hairpins structure. However, a large amount of similar sequences can be folded into this kind of structure. Several existing computational approaches have been developed to predict which hairpins can be pre-miRNAs, but they require a sufficient number of known pre-miRNAs and non pre-miRNAs as learning samples. However, most sequenced genomes have a very small number of miRNAs reported and most of the sequences are unlabeled. The semi-supervised approach proposed in this work takes advantage of these sequences to achieve better prediction rates than state-of-the-art methods [1].

Proposed method

The first step is to build a similarity matrix among the sequences using the euclidean distance between their feature vectors. Then, a vector of labels **y** is defined, having a positive value for known miRNAs, negative for non-miRNAs and zero for unlabeled sequences. Thus, the scores **z** to assign a class to the unlabeled sequences is obtained solving the optimization problem

$$argmin_{\mathbf{z}} \quad \mathbf{z}^T L\, \mathbf{z} + c\,(\mathbf{z}-\mathbf{y})^T C\,(\mathbf{z}-\mathbf{y})$$
$$subject\ to \quad \mathbf{z}\,\mathbf{1} = 0 \quad \text{and} \quad \mathbf{z}^T\mathbf{z} = n$$

where *C* is a diagonal matrix that allows different misclassification cost per sequence, *c* is a regularization parameter and *L* is the normalized Laplacian of the similarity matrix [2]. The first term of the objective function forces similar sequences to have the same labels. The second term penalizes errors in the labeled sequences. The matrix *C* allows to compensate the imbalance in the learning samples, increasing the weights of the minority class (the pre-miRNAs). The first restriction avoids the trivial solution where all sequences have the same label. The second restriction eliminates scaled versions of **z** in the solution space.

Results

To test the prediction power of miRNAss, we have compared it with a similar approach [3] that uses few training examples. Table 1 shows that miRNAss has outperformed it in most cases in the same experiments with human data.

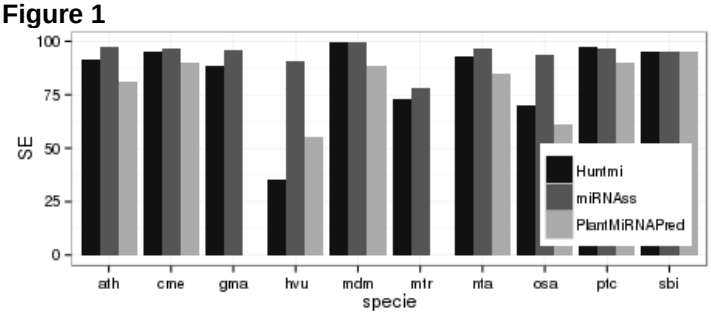
Furthermore, to test miRNAss predictivity in other species, the plant datasets provided by Gudyś *et al*. [4] have been used. In Figure 1 can be seen that, trained with miRNAs from mirBase 17, miRNAss predicted more miRNAs of mirBase 19 than two of the best plant prediction algorithms nowadays: microPlantPred and HuntMi [4].

Table 1

# learning samples	MiRank [3]	miRNAss
1	56.55 %	77.48 %
5	77.85 %	76.41 %
10	79.28 %	82.39 %
15	81.30 %	86.65 %
20	81.48 %	87.42 %
50	82.09 %	89.09 %

Geometric mean of sensibility and specificity with different number of learning samples.

Figure 1



Sensibility for predicting pre-miRNAs included in mirBase 19.

Conclusion

We have presented a new miRNA prediction method called miRNAss. It uses a semi-supervised approach to face the problem of very few training samples within complete genomes. The experiments showed that miRNAss can effectively achieve better results than state-of-the-art methods with very few training samples and that it is versatile enough to be used in genomes of several species.

Reference

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

2. Joachims, T.: **Transductive learning via spectral graph partitioning**. *ICML* 2003, 3:290-297.

3. Xu, Y., Zhou, X., & Zhang, W.: **MicroRNA prediction with a novel ranking algorithm based on random walks**. *Bioinformatics* 2008, ***24*(13)**:i50-i58.

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