High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM

G. Stegmayer, *Member, IEEE,* C. Yones, L. Kamenetzky, D. H. Milone, *Member, IEEE*

Abstract

The computational prediction of novel microRNA within a full genome involves identifying sequences having the highest chance of being a miRNA precursor (pre-miRNA). These sequences are usually named candidates to miRNA. The well-known pre-miRNAs are usually only a few in comparison to the hundreds of thousands of potential candidates to miRNA that have to be analyzed, which makes this task a high class-imbalance classification problem. The classical way of approaching it has been training a binary classifier in a supervised manner, using well-known pre-miRNAs as positive class and artificially defining the negative class. However, although the selection of positive labeled examples is straightforward, it is very difficult to build a set of negative examples in order to obtain a good set of training samples for a supervised method. In this work, we propose a novel and effective way of approaching this problem using machine learning, without the definition of negative examples. The proposal is based on clustering unlabeled sequences of a genome together with well-known miRNA precursors for the organism under study, which allows for the quick identification of the best candidates to miRNA as those sequences clustered with known precursors. Furthermore, we propose a deep model to overcome the problem of having very few positive class labels. They are always maintained in the deep levels as positive class while less likely pre-miRNA sequences are filtered level after level. Our approach has been compared with other methods for pre-miRNAs prediction in several species, showing effective predictivity of novel miRNAs. Additionally, we will show that our approach has a lower training time and allows for a better graphical navegability and interpretation of the results. A web-demo interface to try deepSOM is available at http://fich.unl.edu.ar/sinc/web-demo/deepsom/.

Index Terms

Unsupervised learning, classification, high class-imbalance, deep self-organizing maps.

✦

- *G. Stegmayer, C. Yones and D.H. Milone are with Research Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, CONICET, Argentina (email: gstegmayer@sinc.unl.edu.ar).*
- *L. Kamenetzky is with Instituto de Investigaciones en Microbiologa y Parasitologa Mdica - IMPAM, UBA-CONICET, Argentina.*

1 Introduction

The high class-imbalance problem has been largely recognized as an important issue in machine learning [1], [2] and, more recently, it has been discussed in the context of big data mining [3], [4]. The problem occurs when there are significantly fewer training instances of one class in comparison to another one. Most machine learning algorithms work well with balanced data sets. With imbalanced data sets, however, a supervised classifier can produce a model that tends to be biased towards the majority class and has low performance on the minority one. In fact, the minority class instances are more likely to be misclassified, or even considered noise in some cases [5].

Most of the current standard classification algorithms are designed to maximize the overall number of correct predictions. This criterion is based on an assumption of an equal cost of misclassifications in each class. When the class sizes differ considerably, most standard classifiers would favor the larger class having a high accuracy in prediction (sensitivity if the positive class is the majority one, or specificity if the negative class is the majority class) and the minority class will have a low accuracy. Moreover, it has been studied that for many kinds of classifiers the class imbalance problem is exacerbated when data are high-dimensional since it further increases the bias towards the classification into the majority class, even when there is no real difference between the classes [6]. Supervised classification needs the definition of both positive and negative class samples. Although many proposals have been published on supervised classifiers for imbalanced data sets [7], [6], classification of high class-imbalanced data where one class is significantly under-represented relative to another remains among the leading challenges in the development of prediction models. This is of particular importance in bioinformatics, where there are large biological datasets with this type of unbalanced data.

The presence of high class-imbalance has important consequences on the learning process, usually producing classifiers that have very poor predictive accuracy for the minority class. Thus, this is a main challenge in high class-imbalanced classification nowadays. This problem is of interest in the bioinformatics domain for the computational prediction of microRNAs (miRNAs)[8], where there are only dozens or hundreds (it depends on the organism under study) of well-known miRNAs, versus thousand hundreds of unknown/unlabeled sequences in the rest of the genome, many of which are really negative class and among which there can be hidden candidates to miRNAs.

MicroRNAs are a new type of small RNA molecules, present in both animals and plants, which determine the genetic expression of cells and influence the state of the tissues [9]. Many studies have shown that miRNAs are implied, for example, in cancer progression [10] as well as in viral infection processes [11] and parasites development [12]. Given their role in promoting or inhibiting certain diseases and infections, the discovery of new miRNAs is of high interest today. MicroRNA precursors generated during biogenesis have well-known RNA secondary structures (pre-miRNAs, also known as hairpins) that have allowed the development of computational algorithms for their identification. They typically exhibit a stem-loop structure with few internal loops or asymmetric bulges. However, a large amount of similar hairpins can be folded in many genomes. Due to the difficulty in systematically detecting pre-miRNAs by existing experimental techniques, which have proven to be inefficient and costly, computational methods play an important role nowadays in the identification of new miRNAs [13], [14]. In this context, many computational techniques have emerged lately for identifying miRNAs directly from the characteristics of the RNA sequences. They can be classified into three main categories [13]: i) experimental approaches driven by data, by direct cloning and genome sequencing; ii) comparative methods, based on either sequence or structure conservation between species; and iii) machine learning methods, based on the inherent characteristics (features) of the sequences and secondary structure of these types of molecules. Machine learning methods essentially identify hairpin structures in non-coding and non-repetitive regions of the genome that are characteristic of miRNA precursor sequences. Structures of known miRNAs are used during the learning processes to discriminate between true predictions and false positives [8].

The earliest methods based on machine learning that have been proposed for pre-miRNA identification have used simple representations to extract the main structural features of known pre-miRNAs [15], [16], [17], [18]. For example, their typical stem-loop structure, the frequency of occurrence of nucleotides, the number of base pairs and the minimum free folding energy. It has been established that local sequence features as well as secondary structure are very important for pre-miRNAs identification [19], [20]. However, the definition of the most suitable characteristics to distinguish between true pre-miRNAs and negative cases still remains an important challenge [14]. After the feature extraction step, generally a binary classifier is trained in order to classify or identify sequences highly likely to be miRNA precursors. Support vector machine (SVM) is the learning algorithm that has been most widely applied to solve this problem, using as positive sets the genuine pre-miRNA and artificially defining negative sets of hairpins [15], [21], [22], [23], [24]. Such classification models were expected to perform well in predicting novel pre-miRNAs from unseen sequences. However, although the selection of positive labeled examples for training a binary classifier is really straightforward (known miRNAs), it is very difficult to build a set of negative examples capable of effectively describing this class [19]. A recent study has stated that most of existing machine learning classifiers cannot provide reliable predictive performances on independent testing data sets because the negative training sets are not sufficiently representative [25]. This means that most existing supervised proposals, although reporting very high accuracies, cannot be really trusted in practical and realistic situations.

Methods that use only positive samples to predict new pre-miRNAs, namely one-class classifiers, have been proposed and revised, such as one-class SVM (OC-SVM) and algorithms based on k-nearest neighbors (OC-KNN) [26], [27]. In one-class classifiers, the positive class is learned in a supervised way since positive training data is readily available from miRBase. In contrast, the negative class is not learned at all. Thus, the principal advantage of the one-class approaches is not having to define the negative set for training. However, it has been shown that these models are underperformed in comparison to the two-class approach [27]. The main reason is that one-class methods do not model the negative class, or model it under very simplified assumptions, distant to the real complexity of negative data for this task. Additionally, one-class and two-class methods can provide useful classification accuracies only when there is not very large class-imbalance [28].

In summary, this is one of the main reasons why existing supervised computational methods for new pre-miRNAs identification are not completely satisfactory nowadays [13], [20], [8]. In spite of the fact that many techniques have proven to be a powerful way of distinguishing pre-miRNA hairpins from pseudo hairpins and are implemented in a number of miRNA search tools [29], [30], [31], those methods do not address the high class-imbalance problem properly. This important fact may lead to overlearning the majority class and/or incorrect assessment of classification performance. Moreover, those tools are only effective for a narrow range of species, usually just the model ones such as human, mouse, fly or worm. Given the very large number of candidates to be analyzed (hundreds of thousands sequences), the strong class imbalance between labeled and unlabeled data and the challenge of training with a high percentage of unlabeled data, new strategies must be proposed to address these issues [32].

In this work we present a novel approach for dealing with the high imbalance problem in pre-miRNA prediction. It has been shown that the assumption that many miRNAs occur in clusters can be fruitful for the discovery of novel miRNAs and that most miRNAs often cluster together in portions of the feature space [20]. Thus, instead of training a classifier in a classical supervised manner, we propose to identify miRNA precursors through a novel approach based a hierarchy of self-organizing maps (SOM) organized into a deep architecture (deepSOM), where the best highly-likely candidates to pre-miRNAs (from unlabeled sequences) will be clustered together with well-known pre-miRNAs all along the deeper SOM models. With our proposal, the strong class imbalance problem can be avoided since only positive class examples, even if there are just a few, are necessary. This way, the very-hard to build negative artificial examples must not be defined, making it more useful to analyze genome data from any organism. In fact, the labeling process through wet experiments is very expensive and most of the time, infeasible, whereas acquisition of unlabeled data is relatively inexpensive [8]. During training, the deepSOM classifier is refined level after level, discarding low-quality candidates automatically. Only the best candidates to pre-miRNAs are preserved at each level. At the last level, the sequences assigned to neurons (clusters) that include well-known miRNAs are identified as highly likely candidates to miRNAs. The proposed approach has been tested with several animals and plants miRNAs, using large and varied strongly imbalanced datasets in 10-fold cross-validation tests. As a result, deepSOM has effectively achieved better performance than other existing miRNA prediction tools.

This paper is organized as follows. Section 2 explains the deepSOM architecture and training algorithm in detail. Section 3 presents the data sets used in this study, the experimental setup and performance measures. Section 4 shows the results obtained and their discussion. Finally, the conclusions of this work can be found in Section 5.

2 DEEPSOM FOR HIGH CLASS-IMBALANCED BIOLOGICAL DATA

Self-organizing maps (SOMs) were first introduced in 1982 by Teuvo Kohonen [33]. SOMs are a special class of neural networks that use unsupervised competitive learning, which is based on the idea of units (neurons) that compete to respond to a given set of inputs. Each neuron in a SOM can be considered a cluster, and it is associated with a prototype or synaptic weight vector [34]. Given an input pattern, its distance to the neurons weight vector (centroid or prototype) is computed. Neurons compete with each other, and only the closest neuron prototype to the input becomes activated or fired, becoming this way a winning neuron. The weight vector of this winning neuron is further moved towards the input pattern [35].

The goal of SOM is to represent complex high-dimensional input patterns into a simpler low-dimensional discrete map, with prototype vectors that can be visualized in a two-dimensional lattice structure, while preserving the proximity relationships of the original data as much as possible. Having finished the training, input patterns are projected into the lattice of adjacent neurons, giving a clear topology of how the network fits into the input space. Therefore, the regions with a high probability of occurrence of patterns will be represented by larger areas in the map [36]. That is why SOM can be appropriate for visualization and data analysis when looking for underlying hidden patterns in data. A SOM structures the neurons in a way that those in closer proximity are more similar to each other than to others that are farther apart [37].

In this work, we propose to identify the best candidates to miRNA precursors through a novel machine learning approach based on SOM. This proposal is based on the fact that SOMs have the capability of identifying similar input patterns in the feature space, by assigning them to the same neuron or a group of adjacent neurons on the map [38]. Thus, instead of the classical approach for pre-miRNA prediction that requires training a classifier in a supervised manner with positive and negative classes, in this work we state that only positive examples, together with as many unlabeled sequences as there can be, are necessary for training a SOM for miRNAs prediction. We propose a hierarchy of SOM in deep levels (deepSOM) with the aim of refining the original high level map by discarding unlabeled sequences that are distant to miRNA neurons, level after level.

The training process of deepSOM starts with the root SOM on the first layer. This map undergoes standard training with the complete set of data, using an initial large map size. When this first SOM becomes stable, that is to say, no more further adaptation of the weight vectors occurs, only the data in the neurons having clustered at least one well-known labeled data (plus other many unlabeled sequences) are chosen as input data for training the next map in the second layer. These neurons are denominated miRNA neurons and, although they might contain much more unlabeled data than labeled due to the existing high class-imbalance, they are marked as positive class neurons. The labeling of the miRNA neurons can be done taking into account the neighbouring neurons as well, based on the topologic conservation properties of SOM. That is to say, when a neuron has at least one well-known pre-miRNA, not only this neuron is labeled as miRNA neuron but also its neighbouring neurons (within a certain ratio) can be labeled positive class. During training, only sequences clustered in miRNA neurons remain for further training the next level of deepSOM. After training several nested SOM, the best pre-miRNAs can be identified as the ones that remain close to the prototypes of the miRNA neurons in the last deep level. At the last level, a very small number of sequences (in comparison to the original input size) is the output of the last nested SOM.

Algorithm 1 presents the deepSOM model training and labeling in detail, where the following notation

Algorithm 1: deepSOM training and labeling for pre-miRNA prediction in high class-imbalance data.

```
Inputs :
        $G_\ell$: labeled input sequences
        $G_u$: unlabeled input sequences
        $n$: initial map size ($n \times n$)
        $h_{max}$: maximum deep level
Outputs:
        $\Gamma$: trained neurons at each level
        $\mathcal{L}$: sets of miRNA-neurons for each level
        $C$: pre-miRNA candidates at the last level
1  begin
2      $n_1 \leftarrow n$
3      $D_1 \leftarrow G_u \cup G_\ell$
4      $\mathcal{L} \leftarrow \varnothing$
5      $h \leftarrow 1$
6      while $h < h_{max}$ & $n_h > 1$ & $|D_h| < |D_{h-1}|$ do
7          $\Gamma_h \leftarrow$ Train a $n_h \times n_h$ SOM with $D_h$
8          foreach neuron $i \in \Gamma_h$ do
9              $\gamma_{\Lambda_i} \leftarrow \bigcup_{\forall j \in \Lambda_i} \gamma_j$
10             if $|(\gamma_i \cup \gamma_{\Lambda_i}) \cap G_\ell| > 0$ then
11                 $\mathcal{L}_h \leftarrow \mathcal{L}_h \cup \{i\}$
12         $h \leftarrow h + 1$
13         $D_h \leftarrow \bigcup_{\forall j \in \mathcal{L}_{h-1}} \gamma_j$
14         $n_h \leftarrow \left\lfloor \sqrt{5\sqrt{|D_h|}} \right\rfloor$
15     $C \leftarrow D_{h_{max}}$
```

is used: G_ℓ and G_u are the labeled and unlabeled input training sequences, respectively, where labeled input sequences correspond to well-known miRNAs; n is the initial map size ($n \times n$ neurons); and h is the maximum deep level. The deepSOM training involves the following steps. While the maximum level of deep SOMs has not been reached, and there are data to train a map (line 6), a SOM map is trained at each level (line 7). The top level SOM, at $h = 1$, is set to the initial map size and trained with all input training data (labeled and unlabeled data). During training, each input data point is assigned to a map unit γ_i according to the minimum Euclidean distance between the feature vector representing each sequence and each neuron centroid. Due to the high class-imbalance existing in data, there are only very few positive class labels and most of the data is unlabeled. In large size maps, and because of the topological properties of SOM, neighborhood neurons can be taken into account for neuron labeling, as well. The neighborhood neurons to a given neuron i are indicated as Λ_i and all data samples from neighbors can be taken into account for labeling i (line 9). Neurons labeling occurs by taking into account the labeled data only, as follows: if there is at least one labeled input sequence G_ℓ in the neuron γ_i or within its neighborhood γ_{Λ_i} (line 10), the neuron is labeled as a miRNA-neuron (line 11), no matter how many other unlabeled data points are clustered there as well. Then, only sequences clustered on miRNA-neurons pass to the next deepSOM level (line 13). At each deepSOM level h, the number of neurons $n_h \times n_h$ is determined automatically, according to an heuristics suggested by Kohonen [33], [35], which states that the total number of neurons in a map is related to the number of data points to train it. Thus, n_h is set according to the number of sequences selected in $|D_h|$ (line 14). After training all deepSOM levels, only the data points that are clustered into labeled neurons at the deepest level are predicted as good candidates with a high probability of being miRNA precursors (line 15).

In summary, by training a deepSOM with well-known pre-miRNAs and unlabeled sequences together, knowing that very similar sequences (according to the feature space) will be clustered in the same (or neighboring) neurons, candidate sequences to be real miRNAs can be found by simply inspecting the neurons having well-known pre-miRNA samples at the last level of the deep architecture. The advantages of this proposal are the following. First, only positive label class samples are necessary, no matter if they are just a few in comparison to the many unlabeled ones. Second, the high class-imbalance is being diminished automatically during deepSOM training, level after level of the hierarchy, since the worst (farthest) candidates to miRNAs are filtered in each level and do not pass to the next one; as a consequence the depeest maps have the possibility of better clustering well-known pre-miRNA sequences and unlabeled ones. A third important point is data reduction, because at the last level only the best sequences remain as candidates to pre-miRNAs, no matter how many genome sequences have been used as input to the first SOM model. The most common case in a real application would be hundreds of thousands of sequences, while it is commonly expected that only 10% or less of a genome might contain true miRNAs [8]. Thus, a characteristic that is very desirable in a pre-miRNA classifier: to be able to predict a reasonable number of candidates to be tested in wet experiments, it is actually very hard to provide with a classical supervised model in the presence of high class-imbalance sets. This is another important contribution of deepSOM because the number of candidates requiring further experimental validation is highly reduced within the hierarchy of the maps.

3 Materials and experimental methods

This section describes the datasets used in this work, the experimental setup and the measures used for performance evaluation. The deepSOM source code and training data are freely available for academic purposes and can be found at https://sourceforge.net/projects/sourcesinc/files/deepSOM/. A web-demo [39] interface to rapidly test deepSOM is also available at http://fich.unl.edu.ar/sinc/web-demo/deepsom/.

TABLE 1
Characteristics of the high class-imbalance biological data sets used in the experiments.

Name	Labeled samples	Unlabeled samples	Imbalance ratio
H. sapiens	1406	81228	57.8
A. thaliana	231	28359	122.8
Animals	7053	218154	30.9
Plants	2172	114929	52.9

3.1 High class-imbalance pre-miRNA data sets

The characteristics of the biological data sets used in the experiments are shown in Table 1. For the positive class, all well-known pre-miRNAs in miRBase v17 [40] (except those sequences lacking experimental confirmation) for *Homo sapiens*, *Arabidopsis thaliana*, a set of animals (*Rattus norvegicus, Drosophila melanogaster, Mus musculus, Caenorhabditis elegans, Pan troglodytes, Gallus gallus, Macaca mulatta, Bos taurus, Danio rerio, and Monodelphis domestica*) and plants (*Glycine max, Zea mays, Populus trichocarpa, Selaginella moellendorffii, Triticum aestivum, Vitis vinifera and Oryza sativa*) have been used as in [30].

These data sets include all well-known miRNAs, and in particular the most studied model species *H. sapiens* and *A. thaliana*. Unlabeled sets were built by extracting random sequences from the genomes and mRNAs of these species. The sequence length distribution in the unlabeled 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 220 bases in the stem above 0.15, as in [30]. Class imbalance has been defined as a ratio of number of unlabeled to number of labeled samples. It can be seen from the table that a wide-range of possible imbalance situation have been taken into account, from moderate to very high class-imbalance.

3.2 Experimental setup

For training and testing the deepSOM, a 10-fold cross validation (CV) procedure has been used, giving reliable estimates of classification performance. In all classification experiments, the distributions of testing samples are the same as for the entire datasets. The performance in each experiment is reported as the average values on 10 folds for the test partitions only.

Selecting an informative feature set is very important for the pre-miRNA prediction problem. Most commonly used feature sets contain information about sequence, topology and structure [41]. The earliest machine learning approaches [15] proposed features, named triplets, computed from the sequence itself without including additional characteristics. miPred [42] was the first method that proposed a representative feature set that has shown great discriminative power and that has been adopted by many other current methods [30], [14]. Thus, for fair comparisons with state-of-the-art classifiers, we have used the features of [30] in this study: triplets, maximal length of the amino acid string, cumulative size of internal loops found in the secondary structure, and percentage of low complexity regions detected in the sequence. For each training set, in each fold, an independent 10-fold CV feature selection step has been performed (see details in Supplementary material). In all experiments, deepSOM maximum hidden level has been set to $h = 10$ and initial map size has been set to a large number ($n = 100$). Then, level after level, the map size and labeling neighborhood are automatically determined according to the number of data selected to train the next level SOM.

3.3 Model performance

The prediction quality of the model was assessed by the following classical classification measures: sensitivity (Se), specificity (Sp), accuracy (Acc), and geometric mean (Gm) of classification sensitivity and specificity. These measures are defined as:

$$\begin{aligned} Se &= \frac{TP}{TP+FN}, && (1)\\ Sp &= \frac{TN}{TN+FP}, && (2)\\ Acc &= \frac{TP+TN}{TP+TN+FP+FN}, && (3)\\ Gm &= \sqrt{SE \times SP}, && (4) \end{aligned}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively.

4 Results and discussion

This section presents the results of the experiments made to analyze in detail the behavior of deepSOM for high class-imbalance data sets. After that, comparisons to state-of-the-art miRNA prediction methods are shown for several animals and plants species.

TABLE 2
deepSOM classification results for pre-miRNA prediction in high class-imbalance data sets. Average results are reported on test data in 10-fold CV.

H. sapiens

h	n_h	Se	Sp	Acc	Gm
1	100	0.9722	0.8114	0.8142	0.8881
2	24	0.9722	0.8176	0.8202	0.8915
3	24	0.9698	0.8214	0.8240	0.8925
4	24	0.9680	0.8243	0.8267	0.8932
5	24	0.9664	0.8274	0.8297	0.8941
6	24	0.9606	0.8301	0.8323	0.8929
7	24	0.9592	0.8319	0.8341	0.8932
8	24	0.9564	0.8338	0.8359	0.8930
9	24	0.9564	0.8352	0.8372	0.8937
10	23	0.9556	0.8360	0.8380	0.8937

A. thaliana

h	n_h	Se	Sp	Acc	Gm
1	100	0.9623	0.9554	0.9555	0.9587
2	13	0.9587	0.9576	0.9576	0.9580
3	13	0.9518	0.9599	0.9598	0.9557
4	13	0.9459	0.9614	0.9612	0.9534
5	13	0.9459	0.9624	0.9623	0.954
6	13	0.9396	0.9635	0.9633	0.9512
7	12	0.9360	0.9645	0.9642	0.9499
8	12	0.9360	0.9651	0.9648	0.9501
9	12	0.9360	0.9656	0.9653	0.9504
10	12	0.9304	0.9663	0.9659	0.9479

Animals

h	n_h	Se	Sp	Acc	Gm
1	100	0.9756	0.6502	0.6604	0.7964
2	37	0.9529	0.7123	0.7199	0.8239
3	35	0.9407	0.7372	0.7436	0.8328
4	34	0.9323	0.7540	0.7595	0.8384
5	34	0.9273	0.7675	0.7725	0.8436
6	33	0.9226	0.7767	0.7813	0.8465
7	33	0.9187	0.7839	0.7881	0.8486
8	33	0.9152	0.7909	0.7948	0.8508
9	32	0.9112	0.7982	0.8017	0.8528
10	32	0.9084	0.8045	0.8077	0.8549

Plants

h	n_h	Se	Sp	Acc	Gm
1	100	0.9610	0.7072	0.7119	0.8244
2	29	0.9578	0.7413	0.7453	0.8426
3	29	0.9559	0.7536	0.7573	0.8487
4	28	0.955	0.7597	0.7633	0.8518
5	28	0.9537	0.7664	0.7698	0.8549
6	28	0.9537	0.7709	0.7742	0.8574
7	28	0.9532	0.7751	0.7784	0.8595
8	28	0.9532	0.7783	0.7816	0.8613
9	28	0.9515	0.7815	0.7846	0.8623
10	27	0.9501	0.7839	0.7870	0.8630

4.1 deepSOM in high class-imbalanced data sets

Table 2 shows the deepSOM results for the *Homo sapiens*, *Arabidopsis thaliana*, animals and plants data sets (detailed in Table 1). Average results are reported for test data in 10-fold CV. The first column shows the deepSOM level. The second column shows the map size at each level. From third to sixth column, average Se, Sp, Acc and Gm are reported. This table clearly shows how, in average, very high classification rates are achieved by deepSOM in all cases. For example, with the human data set, the deepSOM performance for recognizing human pre-miRNAs is very high at the top SOM (97.22%). At deeper levels, the Se is slightly worsened (up to 96% approximately) and at the same time that the model improves Sp rate, that is to say, better discarding not-good candidates to pre-miRNAs. The detail of the Acc and Gm evolution shows how deepSOM is being refined level after level, achieving better and better global performance at deeper levels. In fact, overall, for human data the deepSOM after 10 levels has achieved a very good Gm of almost 90%. It has to be taken into account that for this data set, in each testing fold, the number of sequences presented to the top-level deepSOM is around 8000 sequences, remaining only 1700 candidates at the last level, in average. Thus, the reduction in the number of candidates at the last level is worth to be highlighted. This shows how, at the same time that the performance measures are mantained at high values in deeper levels and even improved (as in the case of Acc and Gm), the number of candidates to pre-miRNA that survive level after level is dramatically reduced after the first step, and subsequently refined.

In the other data sets, the same general behavior regarding Se, Sp, Acc and Gm can be observed. In all cases, very high Se values, between 91% and 98%, are reached, depending on the species. Even for the most imbalanced data set (*A. thaliana*) as well as in the two largest ones, animal and plants, the Se values are higher than 90%. In all cases, the Sp values improves with more hidden levels. The global performance measures, Acc and Gm, improve also in all cases level after level. In this table, it can be clearly seen that in general, and for all data sets, their values are being increased level after level, maintaining a constant value at the depeest levels. Regarding the most imbalanced data set, it is worth to highlight the fact that deepSOM reaches almost 95% of Gm and 97% of accuracy.

The training time is an important issue determining the applicability of the deepSOM method to real-life problems. Table 3 shows the performance of deepSOM for miRNA prediction in miRBase v17. Training time is reported as median over 10 cross-validation training folds in format *hh:mm:ss*. The table shows that deepSOM training and labeling is always in the range of minutes, around 10 minutes, for all data sets, even the largest ones. For these same data sets, in comparison, training time in the range of several hours have been reported in [30] due to the fact that an exhaustive parameter search is performed. Our proposal, instead, even for the most high class-imbalanced and large sets, has an speed of training and execution extremely fast: deepSOM is more than 10 times faster than the cited work for the most imbalanced dataset. Regarding the largest data sets, deepSOM achieves running times in the order of minutes also, against several hours of computation reported in [30].

Table 4 shows the number of sequences that remain clustered in miRNA neurons at each corresponding level in deepSOM, for each studied data set. Level $h = 0$ indicates the total number of sequences that are input to a SOM for pre-miRNA prediction. From level $h = 1$ to $h = 10$, it is indicated the number of data samples that remained clustered in miRNA neurons and, therefore, pass from one level to the next one for training another SOM. It can be clearly seen here the significant reduction in the number

TABLE 3
Training times of deepSOM for pre-miRNA prediction in miRBase 17. Median over all 10-fold cross-validation training partitions are reported in format *hh:mm:ss*.

Dataset	training time
H. sapiens	00:07:43
A. thaliana	00:04:34
Animals	00:11:21
Plants	00:06:03

TABLE 4
deepSOM results for pre-miRNA prediction in high class-imbalance data sets: number of pre-miRNA candidates as input in each level.

h	*H. sapiens*	*A. thaliana*	Animals	Plants
0	82634	28590	225207	117101
1	2342	1542	85883	37986
2	2224	1483	84439	33593
3	2043	1363	83978	31707
4	1971	1276	83827	30795
5	1928	1266	83670	30088
6	1927	1257	83593	29405
7	1918	1233	83468	28801
8	1885	1229	83402	28626
9	1868	1159	83039	28183
10	1853	1159	82399	27714

of pre-miRNA candidates, level after level. The high class-imbalance is greatly diminished after the first large map, and the successive levels refine the deepSOM prediction model, until it reaches a level where the number of candidates that pass from one map to the next one almost does not change.

An additional way of viewing this data reduction is through maps visualization at each level. This is an additional feature of the deepSOM, which helps to the interpretability of the results. The projection of training data into the deepSOM lattice of neurons at each level gives a clear view of the topological distribution of well-known pre-miRNAs (labeled class) with respect to unlabeled samples. As an example, Figure 1 shows the 10 deep maps corresponding to the *Arabidopsis thaliana* data set. In the figure, the left map shows all input data projected into the deepSOM model at $h = 1$, where this top level map size is 100x100 neurons. The right maps, from up-to-down and left-to-right, show the following 2 to 10 maps. The neurons that have well-known miRNAs are painted in red. The neurons that have unlabeled sequences (and will be discarded in the next level) are painted in blue. The neurons that have labeled and unlabeled samples (mixed neurons) are indicated with gray. The marker size indicates the number of

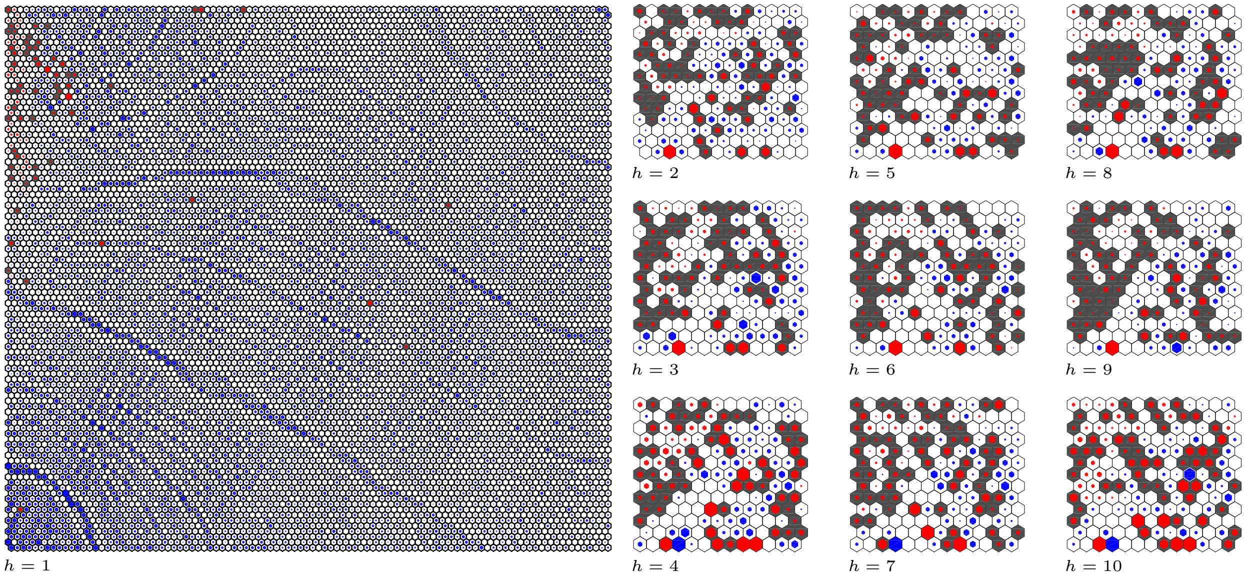


Fig. 1. Example of deepSOM model for the *A. thaliana* data: miRNA neurons (red), no-miRNA neurons (blue), mixed neurons (gray). Marker size indicates number of sequences in each neuron.

samples that is clustered in each neuron.

This visualization of the deep maps obtained for a dataset shows, very quickly, how the pre-miRNAs are clustered nearby into a portion of the feature map (in the example, at the top-left corner of the first big map); while a very large amount of unlabeled sequences are clustered in the rest of the map, clearly away form the labeled class. This large amount of data is discarded and not passed to the second level deepSOM. Thus, in this first step, there is an important data reduction and a significant high class-imbalance reduction as well, thanks to the topological properties of SOM. The use of different markers size helps understanding the data magnitudes involved in the problem. The subsequent maps (from $h = 2$ to $h = 10$) are very similar among them because they have already received, from the first map, the best candidates to miRNAs and now they have to just specialize more for detecting a reduced list of best pre-miRNAs. In fact their sizes practically do not change (see Table 2) in the last three levels and more empty neurons appear because the most similar samples are clustered together, more accurately, while unlabeled data to be discarded (in the blue neurons) is being reduced. This is a very useful feature of the proposal because it helps having a real idea of the number of trully candidates inside a genome, their distribution into the feature space according to the features chosen for their representation, and also, the existing data class imbalance can be actually seen in a clear and simple graphical way. This visualization could be used to explore the feature space distribution of a genome data: several deepSOM models could be built, using different feature sets, and compared in order to see which map could be better for pre-miRNAs

TABLE 5
Comparison with other tools: animal species. Classification sensitivity of deepSOM on animal miRNAs recently added in miRBase v18-19.

Species	MicroPred [42]	OC-KNN [26]	OC-SVM [27]	HuntMi [30]	deepSOM
Bombyx mori	75.00	50.00	50.00	**100.00**	**100.00**
Caenorhabditis elegans	87.50	31.25	50.00	93.75	**100.00**
Ciona intestinalis	89.47	63.16	52.63	73.68	**94.74**
Homo sapiens	85.14	52.00	33.71	93.14	**96.00**
Macaca mulatta	-	56.25	37.50	81.25	**87.50**
Mus musculus	64.03	58.27	46.04	**94.96**	**94.96**
Oryzias latipes	94.08	57.89	60.53	96.05	**97.37**
Pongo pygmaeus	83.33	42.59	48.15	**94.44**	**94.44**
Rattus norvegicus	76.32	55.26	60.53	**97.37**	**97.37**
Taeniopygia guttata	82.61	34.78	26.09	91.30	**100.00**
Tribolium castaneum	64.29	64.29	50.00	78.57	**100.00**

TABLE 6
Comparison with other tools: animal species. Classification sensitivity of deepSOM on animal miRNAs recently added in miRBase v20-21.

Species	sequences	OC-KNN [26]	OC-SVM [27]	HuntMi [30]	deepSOM
Bombyx mori	2	50.00	50.00	**100.00**	**100.00**
Caenorhabditis elegans	50	50.00	42.00	89.80	**100.00**
Ciona intestinalis	19	63.16	52.63	72.22	**94.74**
Homo sapiens	467	52.68	44.11	89.91	**95.72**
Macaca mulatta	140	54.29	31.43	92.81	**96.43**
Mus musculus	492	51.02	48.98	91.04	**95.12**
Oryzias latipes	152	57.24	59.87	96.69	**98.03**
Pongo pygmaeus	63	41.27	50.79	**95.16**	93.65
Rattus norvegicus	87	64.37	49.43	97.67	**98.85**
Taeniopygia guttata	14	33.33	25.93	76.92	**96.30**
Tribolium castaneum	27	64.29	50.00	88.46	**100.00**

prediction. For example, a top-level map with very disperse labeled class samples all over the map might not be preferable over a map that has the known miRNAs samples clustered nearby in a specific zone of the map. Finally, it can be highlighted that for the particular example shown in this figure, the top level deepSOM has received 28,590 sequences, indicating just about 1000 sequences as the highly likely candidates to pre-miRNAs.

4.2 Comparison with other tools: animal species

To further test the performance of deepSOM in a realistic scenario, we trained it on the entire animal dataset from miRBase v17 and tested with animal miRNAs newly introduced in miRBase v18-19 (a test set of 206 sequences), and with the newest release v20-21 (a test set of 1513 new sequences). We compared the performance obtained with deepSOM for this task against two recently proposed miRNA prediction tools, HuntMi [30] and MicroPred [42], as well as versus the one-class classifiers, OC-KNN [26] and OC-SVM [27]. MicroPred has proven to be the best software for human miRNA prediction at the time of its publication; thus its predecessors such as Triplet-SVM [15] or MiPred [17] have not been considered in the comparisons. The obtained results are shown in Tables 5 and 6. Results not reported have not been found in the original work.

The tables clearly demonstrate that deepSOM is capable of efficiently identifying novel microRNAs in animals, achieving a sensitivity of over 90% in 10 out of 11 analysed species. Furthermore, the proposed model has clearly outperformed other state of the art classifiers in most species and equaling the performance of a very recent proposal, HuntMi, in four species. DeepSOM has achieved even 100% effective recognition in four out of eleven cases.

It is worth highlighting here that, although MicroPred is a tool that has been designed specifically for human pre-miRNAs prediction, has achieved worst prediction rates than deepSOM in human miRNAS, and in all other animal test species as well. For the specific test case with newly discovered human miRNAs added to miRBase v18-19, deepSOM has achieved a very high prediction rate (96.00%, 168 out of 175) against 85.14% of MicroPred. HuntMi, in this particular test case, recognised 93.4% of new human miRNAs. This is a quite significant improvement of deepSOM over state-of-the-art methods for the discovery of new miRNAs in such well-studied genome as it is *H. Sapiens*. In the test with the most recent version of miRBase (Table 6), deepSOM has also shown a very high sensitivity rate, having even a better performance than the v18-19 test for some species, such as *H. Sapiens* and *R. Norvegicus*.

In this kind of tests, it is a very important issue to provide measures about the true negative rate of the model, because it is hard to measure it in tests where the only well-known samples are the positive ones. In order to calculate some sort of specificity for the deepSOM model in this real prediction task, and to better illustrate the predictive performance of the proposed approach on completely independent test data, the following two experiments have been done. First, we have re-trained the deepSOM model with the full animals dataset built upon miRBase v17, but leaving now a completely separate set of negative data sequences for test (10%, randomly selected). In this experiment, the specificity of the deepSOM model was 80.70%. Additionally, we have tested a trained deepSOM model with a completely independent negative test set composed by 1,000 human pre-miRNAs, the one defined by the pioneer work of [15]. The specificity of the deepSOM model in this test was of 81.55%, while the HuntMi specificity for this same test was 72.37%.

In summary, all of these tests show that deepSOM can be reliably used for predicting whether sequences in an animal genome can be pre-miRNAs or not with very high confidence.

TABLE 7
Comparison with other tools: plant species. Classification sensitivity of deepSOM on plants miRNAs recently added in miRBase v18-19.

Species	PlantMiRNAPred [43]	OC-KNN [26]	OC-SVM [27]	HuntMi [30]	deepSOM
Arabidopsis thaliana	80.88	54.41	35.29	91.18	**92.65**
Cucumis melo	90.00	41.67	64.17	**95.00**	**95.00**
Glycine max	-	52.65	55.63	88.41	**94.70**
Hordeum vulgare	55.56	46.67	40.00	35.56	**82.22**
Malus domestica	88.83	64.08	57.77	99.51	**100.00**
Medicago truncatula	-	60.33	34.33	72.67	**87.33**
Nicotiana tabacum	84.66	50.92	61.35	93.25	**94.48**
Oryza sativa	60.95	49.11	42.01	69.82	**80.47**
Populus trichocarpa	89.89	48.31	56.18	97.75	**98.88**
Sorghum bicolor	94.83	55.17	37.93	94.83	**100.00**

TABLE 8
Comparison with other tools: plant species. Classification sensitivity of deepSOM on plants miRNAs recently added in miRBase v20-21.

Species	sequences	OC-KNN [26]	OC-SVM [27]	HuntMi [30]	deepSOM
Arabidopsis thaliana	95	56.84	34.74	89.36	**91.58**
Cucumis melo	120	41.67	64.17	94.96	**95.00**
Glycine max	370	51.89	52.43	90.51	**95.95**
Hordeum vulgare	48	43.75	37.50	38.30	**85.42**
Malus domestica	205	64.39	58.05	99.51	**100.00**
Medicago truncatula	327	59.63	34.56	74.85	**88.38**
Nicotiana tabacum	162	51.23	61.11	93.79	**94.44**
Oryza sativa	179	48.04	41.90	71.35	**81.56**
Populus trichocarpa	129	45.74	56.59	**98.44**	96.90
Sorghum bicolor	57	54.39	36.84	94.64	**100.00**

4.3 Comparison with other tools: plant species

To further evaluate the performance of deepSOM for plant miRNA prediction, we have trained it on the full plant dataset built with miRBase v17, and tested it on plant miRNAs introduced in miRBase v18-19 (test set with 1520 sequences) and in the most recent miRBase v20-21 (test set with 1647 sequences). The comparative results of deepSOM on this task against HuntMi [30], one of the most recent methods specialising in plant microRNA identification, PlantMiRNAPred [43], and one-class classifiers OC-KNN [26] and OC-SVM [27]. Comparative results are presented in Tables 7 and 8.

Clearly, deepSOM is superior to the other state-of-the-art methods for miRNA prediction in plants. Only in one case it has achieved the same performance of the most recent proposed classifier. Regarding the comparison with the specific plant pre-miRNA classifer, it is worth highlighting that deepSOM has outperformed it in all test cases. A particular group of plant pre-miRNAs, formed by *H. vulgare*, *M. truncatula* and *O. sativa*, was characterised by very low sensitivity values in the case of existing methods. The authors of HuntMi looked into these pre-miRNAs in detail and discovered that a large fraction of miRNAs in these species do not meet commonly recognised criteria for annotation of plant miRNAs, while in some other miRNAs the mature microRNA lies outside the stem part of the hairpin. This makes more difficult the prediction of miRNAs for these species. In spite of this specific fact, deepSOM achieved very satisfactory recall rates in all of these cases, increasing the sensitivity for the hardest case (*H. vulgare*, from 55.56% [43] and 35.56% [30] to 82.22% for deepSOM). In the more updated test (miRBase v20-21, Table 8), deepSOM has even increased its average sensitivity rate.Similarly to the animal model, in order to show that deepSOM is capable of correctly identifying positive miRNAs and rejecting false miRNAs, we have re-trained the deepSOM model with the full plants dataset built upon miRBase v17, but leaving out a set of negative data sequences for test (10%, randomly selected), achieving a specificity of 79.56%.

In summary, the model proposed in this work has achieved the best recognition rates in all test cases. In this plant data set, deepSOM has achieved a very high classification recall in most cases, higher than 90% in seven out of 10 test species, and even reaching a proportion of correctly identified miRNAs of 100% for two species.

5 Conclusions

In this work we have presented a new and effective approach for the computational prediction of novel microRNAs precursors. As opposite to the classical supervised classifiers generally used for this problem, it does not require the artificial and costly definition of a negative class for training. The proposal involves clustering well-known pre-miRNAs together with unlabeled sequences. This way, clusters having both known miRNAs and other sequences allow the quick identification of the best candidates to be novel pre-miRNAs. The proposed approach, named deepSOM, deals with the high class-imbalance problem in a data set having very few known positive class samples and an excessively larger number of unlabeled sequences through the model organization into a hierarchical architecture of several deep maps. The use of a hierarchy of deep SOM models overcomes the problem of having very few positive class labels, since they are always maintained in the deeper levels, filtering less likely pre-miRNA sequences.

The deepSOM has been tested with several class imbalance real biological data sets, having different levels of imbalance, showing high accuracy results in all cases. The deepSOM performance has been further compared with other state-of-the-art methods for the prediction of novel miRNAs in animals and plants, showing better performance in all tests, even for many different and difficult species. Additionally, we have shown that the proposal allows for a better graphical interpretation of the results when input data is projected graphically into each deep map. The painting of the neurons in the map with different colors and marker sizes according to the type of data clustered, helps having a clear view of the actual high class-imbalance in the problem under study and the candidates to miRNAs distribution and location in the feature space.

Acknowledgements

This work was supported by National Scientific and Technical Research Council [PIP 2013 117], National University of Litoral [CAI+D 2011 548] and Agencia Nacional de Promocion Cientifica y Tecnologica (ANPCyT) [PICT 2014 2627].

References

[1] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, “On the class imbalance problem,” in *Natural Computation, 2008. ICNC ’08. Fourth International Conference on*, vol. 4, Oct 2008, pp. 192–201.

[2] S. Wang and X. Yao, “Multiclass imbalance problems: Analysis and potential solutions,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 1119–1130, Aug 2012.

[3] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[4] C. L. P. Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data.” *Information Sciences*, pp. 314–347, 2014.

[5] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, Eds., *Learning Pattern Classification Tasks with Imbalanced Data Sets, in Pattern Recognition*. InTech, 2009.

[6] R. Blagus and L. Lusa, “Smote for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, no. 1, p. 106, 2013.

[7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 4, pp. 463–484, July 2012.

[8] B. Liu, J. Li, and M. Cairns, “Identifying mirnas, targets and functions,” *Briefings in Bioinformatics*, vol. 15, no. 1, pp. 1–19, 2014.

[9] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function.” *Cell*, vol. 116, pp. 281–297, 2004.

[10] A. Esquela-Kerscher and F. J. Slack, “Oncomirs - microRNAs with a role in cancer,” *Nature Reviews Cancer*, vol. 6, no. 1, pp. 259–269, 2006.

[11] C.-H. Lecellier, P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, C. Himber, A. Saib, and O. Voinnet, “A cellular MicroRNA mediates antiviral defense in human cells,” *Science*, vol. 308, no. 5721, pp. 557–560, 2005.

[12] M. Rosenzvit, M. Cucher, L. Kamenetzky, N. Macchiaroli, L. Prada, and F. Camicia, *MicroRNAs in Endoparasites*. Nova Science Publishers, 2013.

[13] L. Li, J. Xu, D. Yang, X. Tan, and H. Wang, “Computational approaches for microRNA studies: a review,” *Mamm Genome*, vol. 21, no. 1, pp. 1–12, 2010.

[14] Ivani de ON Lopes and Alexander Schliep and Andre de Carvalho, “The discriminant power of RNA features for pre-miRNA recognition,” *BMC Bioinformatics*, vol. 15, no. 1, pp. 124+, 2014.

[15] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, no. 1, p. 310, 2005.

[16] S. A. Helvik, O. Snove, and P. Saetrom, "Reliable prediction of Drosha processing sites improves microRNA gene prediction." *Bioinformatics*, vol. 23, no. 2, 2007.

[17] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Research*, vol. 35, no. 1, pp. W339–W344, 2007.

[18] K. Gkirtzou, I. Tsamardinos, P. Tsakalides, and P. Poirazi, "MatureBayes: A probabilistic algorithm for identifying the mature miRNA within novel precursors," *PLOS one*, vol. 5, no. 8, p. e11843, 2010.

[19] Y. Xu, X. Zhou, and W. Zhang, "MicroRNA prediction with a novel ranking algorithm based on random walks," *Bioinformatics*, vol. 24, no. 1, pp. i50–i58, 2008.

[20] N. D. Mendes, S. Heyne, A. T. Freitas, M.-F. Sagot, and R. Backofen, "Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches," *Bioinformatics*, vol. 28, no. 23, pp. 3034–3041, 2012.

[21] J. Hertel and P. F. Stadler, "Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data," *Bioinformatics*, vol. 22, no. 14, pp. e197–e202, 2006.

[22] T. H. Huang, B. Fan, M. Rothschild, Z. L. Hu, K. Li, and S. H. Zhao, "MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans," *BMC Bioinformatics*, vol. 8, no. 1, pp. 341+, 2007.

[23] J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features," *BMC Bioinformatics*, vol. 11, no. 11, p. S11, 2010.

[24] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, and S. Mavroudi, "YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, p. doi:10.1109/TCBB.2014.2388227, 2015.

[25] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 1, pp. 192–201, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2013.146

[26] M. Yousef, S. Jung, and M. Showe, "Learning from positive examples when the negative class is undetermined- microRNA gene identification," *Algorithms for Molecular Biology*, vol. 3, no. 1, pp. 2–10, 2008.

[27] M. Yousef, N. Najami, and W. Khalifa, "A comparison study between one-class and two-class machine learning for microRNA target detection," *J. Biomedical Science and Engineering*, vol. 3, no. 1, pp. 247–252, 2010.

[28] C. Gomes and et al., "A review of computational tools in microRNA discovery," *Frontiers in Genetics*, vol. 4, no. 1, pp. 81–104, 2013.

[29] D. Gao, R. Middleton, J. Rasko, and W. Ritchie, "miREval 2.0: a web tool for simple microRNA prediction in genome sequences," *Bioinformatics*, vol. 29, no. 24, pp. 3225–3226, 2013.

[30] A. Gudy, M. Szczeniak, M. Sikora, and I. Makalowska, "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification," *BMC Bioinformatics*, vol. 14, no. 1, pp. 83+, 2013. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-14-83

[31] J. An, J. Lai, A. Sajjanhar, M. L. Lehman, and C. C. Nelson, "mirplant: an integrated tool for identification of plant mirna from RNA sequencing data," *BMC Bioinformatics*, vol. 15, p. 275, 2014.

[32] S. Dua and P. Chowriappa, Eds., *Data Mining for bioinformatics*. CRC Press, 2012.

[33] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[34] R. Xu and D. C. Wunsch, *Clustering*. Wiley and IEEE Press, 2009.

[35] T. Kohonen, M. R. Schroeder, and T. S. Huang, *Self-Organizing Maps*. Springer-Verlag New York, Inc., 2005.

[36] S. Haykin, *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Prentice-Hall, Inc., 2007.

[37] G. Stegmayer, M. Gerard, and D. Milone, "Data mining over biological datasets: an integrated approach based on computational intelligence," *IEEE Computational Intelligence Magazine, Special Issue on Computational Intelligence in Bioinformatics*, vol. 7, no. 4, pp. 22–34, 2012.

[38] D. Milone, G. Stegmayer, L. Kamenetzky, M. López, J. Lee, J. Giovannoni, and F. Carrari, "*omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants," *BMC Bioinformatics*, vol. 11, pp. 438–447, 2010.

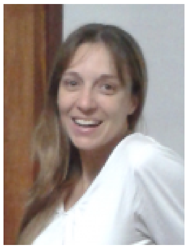
[39] G. Stegmayer, M. Pividori, and D. Milone, "A very simple and fast way to access and validate algorithms in reproducible research," *Briefings in Bioinformatics*, Advance Access published July 28, 2015, doi: 10.1093/bib/bbv054.

[40] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, pp. 152–157, 2011.

[41] C. Yones, G. Stegmayer, L. Kamenetzky, and D. Milone, "miRNAfe: a comprehensive tool for feature extraction in microRNA prediction," *BioSystems*, vol. 238, pp. 1–5, 2015.

[42] R. Batuwita and V. Palade, "*microPred*: effective classification of pre-mirnas for human mirna gene prediction," *Bioinformatics*, vol. 25, no. 8, pp. 989–995, 2009.

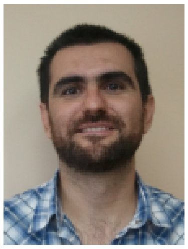
[43] P. Xuan, M. Guo, X. Liu, Y. Huang, W. Li, and Y. Huang, "*PlantMiRNAPred*: efficient classification of real and pseudo plant pre-mirnas," *Bioinformatics*, vol. 27, no. 10, pp. 1368–1376, 2011.



Georgina Stegmayer received the Engineering degree in Information Systems from UTN-FRSF, Argentina, in 2000, and the Ph.D. degree from Politecnico di Torino, Italy, in 2006. Since 2007 she is Assistant Professor of Artificial Intelligence and Computationl Intelligence in UNL University in Argentina. She is currently Adjunct Researcher at the National Scientific and Technical Research Council (CONICET) of Argentina. She is author and co-author of numerous papers on journals, book chapters and conference proceedings on artificial neural networks for a wide variety of problems. Her current research interests involve machine learning, data mining and pattern recognition in bioinformatics.

Cristian Yones received the Computer Engineering degree in 2014 from National University of Litoral (UNL), Argentina. He receive a doctoral scholarship from National Council of Scientific and Technical Research (CONICET) and since 2014 he is a PhD student in Computational Intelligence, Signals and Systems. His research interests include machine learning, data-mining, semi-supervised learning, with applications in bioinformatics.

Laura Kamenetzky received the Master Degree in Biology in 2001 and the PhD in Biological Sciences from Buenos Aires University (UBA), Argentina in 2006. Since 2008 she is Researcher at National Council of Scientific and Technological Research (CONICET) and Assistant Professor of Genetics and Molecular Biology at UBA, Argentina. She is author and co-author of numerous papers on journals, book chapters and conference proceedings on Molecular Biology, Genomics and Bioinformatics. Her current research interests involve the large-scale genome and transcriptome data integration with phenotyping to understand the biology of parasitic flatworm species responsible for human diseases.



Diego H. Milone received the Bioengineering degree (Hons.) from National University of Entre Rios (UNER), Argentina, in 1998, and the Ph.D. degree in Microelectronics and Computer Architectures from Granada University, Spain, in 2003. He was with the Department of Bioengineering and the Department of Mathematics and Informatics at UNER from 1995 to 2002. Since 2003 he is Full Professor in the Department of Informatics at National University of Litoral (UNL). From 2009 to 2011 was Director of the Department of Informatics and from 2010 to 2014 was Assistant Dean for Science and Technology. Since 2006 he is a Research Scientist at the National Scientific and Technical Research Council (CONICET). Since 2015 he is Director of the Research Institute for Signals, Systems and Computational Intelligence (CONICET-UNL). His research interests include statistical learning, pattern recognition, signal processing, neural and evolutionary computing, with applications to speech recognition, affective computing, biomedical signals and bioinformatics.

High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM **Supplementary Material**

G. Stegmayer, C. Yones, L. Kamenetzky, D. H. Milone

Selecting an informative feature set is very important for the pre-miRNA prediction problem. Most commonly used feature sets contain information about sequence, topology and structure. The earliest machine learning approaches proposed a feature named triplets, computed from the sequence itself without including additional characteristics. miPred was the first method that proposed a representative feature set that has shown great discriminative power and that has been adopted by many other current methods, such as PlantmiRNAPred and HuntMi.

For fair comparisons with those state-of-the-art classifiers, we have used those same features in this study:

- G+C content: calculated as (G + C)/(G + C + A + U);
- MFEI1: ratio between the minimum free energy (MFE obtained with the algorithm from [1]), and the G+C content;
- MFEI2: ratio between the dG and the number of stems;
- MFEI3: ratio between the dG and number of loops;
- MFEI4: ratio between the dG and the G+C content;
- dG: adjusted MFE. MFE divided by the sequence length;
- dQ: adjusted Shannon entropy, which characterizes the probability of base pairing in a secondary structure as a chaotic dynamic system;
- dF: measures the compactness of a tree-graph where each vertex represents a bulge loop, hairpin loop, internal loop, the 5' and 3' unpaired ends, or the multi-branch loop and each edge is a RNA stem;
- zD: standard score of the base pair distance. Adjusted base pair distance normalized using z-score.
- Diversity: set diversity obtained with the algorithm from [2];
- NEFE: normalized ensemble free energy;
- Diff: difference between MFE and EFE;
- dS: structure entropy;
- dS/L: normalized structure entropy;
- $|A-U|/L$: number of base pairs A-U normalized with the length;
- $|G-C|/L$: number of base pairs G-C normalized with the length;
- $|G-U|/L$: number of base pairs G-U normalized with the length;
- Avg_BP_Stem: average base pair (nucleotides) per stem (a structural motif of the secondary structure that has more than three contiguous base pairs)
- %(A-U)/n_stem: base pair proportion A-U per stem;
- %(G-C)/n_stem: base pair proportion G-C per stem;
- %(G-U)/n_stem: base pair proportion G-U per stem;
- triplets: frequencies of the following secondary structure triplets composed of three adjacent nucleotides and the middle nucleotide: ”A(((”, ”U(((”, ”G(((”, and ”C(((”;
- maximal length of the amino acid string without stop codons found in three reading frames;
- cumulative size of internal loops found in the secondary structure;
- a percentage of low complexity regions detected in the sequence using Dustmasker [3].

For each training set, in each fold, an independent 10-fold cross-validation (stratified) feature selection step has been performed by using WEKA[1] CfsSubsetEval algorithm [4], which evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

The final set of features that have been selected by dataset are:

Animals:
- MFEI1
- dQ
- Avg_Bp_Stem
- MFEI3
- loops

Plants:
- MFEI1
- dG
- dQ
- Avg_Bp_Stem
- MFEI3
- loops

H. sapiens:
- MFEI1
- dQ
- Avg_Bp_Stem
- MFEI3
- loops

A. thaliana:
- MFEI1
- dG
- Avg_Bp_Stem
- MFEI3
- loops

Tables 1 and 2 report the classification performance of deepSOM on pre-miRNAs recently added in miRBase (v20-21), with the comparison of full features versus feature selection, for the animals and plants datasets, respectively.

References

[1] M. Zuker and P. Stiegler, "Optimal computer folding of large rna sequences using thermodynamics and auxiliary information," *Nucleic Acids Research*, vol. 9, no. 1, pp. 133–148, 1981.

[2] J. McCaskill, "The equilibrium partition function and base pair probabilities for RNA secondary structure," *Biopolymers*, vol. 29, pp. 1105–1119, 1990.

[3] A. Morgulis, E. Gertz, A. Schaffer, and A. R., "A fast and symmetric DUST implementation to mask low-complexity DNA sequences," *Journal of Computational Biology*, vol. 13, pp. 1028–1040, 2006.

[4] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Department of Computer Science, The University of Waikato, 1999.

TABLE 1
Classification performance of deepSOM on animal pre-miRNAs recently added in miRBase (v20-21)

	Full	Selected
Species	features	features
Bombyx mori	100.00	100.00
Caenorhabditis elegans	93.75	100.00
Ciona intestinalis	78.95	94.74
Homo sapiens	98.86	96.00
Macaca mulatta	100.00	87.50
Mus musculus	97.84	94.96
Oryzias latipes	99.34	97.37
Pongo pygmaeus	98.15	94.44
Rattus norvegicus	100.00	97.37
Taeniopygia guttata	95.65	100.00
Tribolium castaneum	100.00	100.00

TABLE 2
Classification performance of deepSOM on plants pre-miRNAs recently added in miRBase (v20-21)

	Full	Selected
Species	features	features
Arabidopsis thaliana	89.71	92.65
Cucumis melo	97.50	95.00
Glycine max	92.05	94.70
Hordeum vulgare	71.11	82.22
Malus domestica	100.00	100.00
Medicago truncatula	80.67	87.33
Nicotiana tabacum	94.48	94.48
Oryza sativa	82.84	80.47
Populus trichocarpa	94.38	98.88
Sorghum bicolor	94.83	100.00

[1] http://www.cs.waikato.ac.nz/ml/weka/